MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL INFORMATION PROCESSING
WHITAKER COLLEGE

A.I. Memo No. 1038 April 1988
C.B.I.P. Memo No. 32

The Computational Study of Vision

Ellen C. Hildreth and Shimon Ullman

Abstract: Through vision, we derive a rich understanding of what is in the world, where objects are located, and how they are changing with time. Because we obtain this understanding immediately, effortlessly, and without conscious introspection, we can be deceived into thinking that vision should therefore be a fairly simple task to perform. The computational approach to the study of vision inquires directly into the sort of information processing needed to extract important information from the changing visual image – information such as the three-dimensional (3-D) structure and movement of objects in the scene, or the color and texture of object surfaces. An important contribution that computational studies have made is to show how difficult vision is to perform, and how complex are the processes needed to perform visual tasks successfully. This article reviews some computational studies of vision, focusing on edge detection, binocular stereo, motion analysis, intermediate vision and object recognition.

Some of the research described in this article was done within the Artificial Intelligence Laboratory and the Center for Biological Information Processing (Whitaker College) at the Massachusetts Institute of Technology. Support for the Artificial Intelligence Laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-85-K-0124. Support for this research is also provided by the Sloan Foundation, the Office of Naval Research, Cognitive and Neural Systems Division, the National Science Foundation (IRI-8657824) and the McDonnell Foundation.

1. INTRODUCTION

Through vision, we derive a rich understanding of what is in the world, where objects are located, and how they are changing with time. Because we obtain this understanding immediately, effortlessly, and without conscious introspection, we can be deceived into thinking that vision should therefore be a fairly simple task to perform. The computational approach to the study of vision inquires directly into the sort of information processing needed to extract important information from the changing visual image — information such as the three–dimensional (3–D) structure and movement of objects in the scene, or the color and texture of object surfaces. An important contribution that computational studies have made is to show how difficult vision is to perform, and how complex are the processes needed to perform visual tasks successfully.

Levels of analysis

The development of computers with increasing power and sophistication has often stimulated comparisons between computers and the human brain, especially since computers have been applied more and more to tasks that were formerly considered uniquely human capabilities, such as understanding natural language. It is clear, however, that at the level of their hardware, neurons and computer circuits are very different. We can nevertheless attempt to describe the processes that take place in these two systems at a level that is essentially independent of this hardware — this is a description of the tasks that they perform. In much the same way that we can describe the theory of arithmetic independent of the computing device that carries out the arithmetic operations, we can describe the theory of vision independent of the hardware that carries it out, whether it be biological or computer hardware.

This idea of separating the tasks performed by a vision system from the hardware that carries out these tasks was central to the work of David Marr (1982). Marr argued that there are at least three different levels at which problems in vision can be described, which he labelled the computational theory, algorithm, and mechanism. Theoretical issues include an analysis of the way in which properties of the physical world constrain how problems in vision are solved. An algorithm is a step–by–step procedure that transforms one representation of visual information into the next. Finally, the mechanism refers to the details of how a computation is carried out in neural or computer hardware. These three levels of analysis are a useful methodological distinction for studying visual processing, but the distinction between levels such as algorithm and mechanism is often not a sharp one for the brain.

An aspect of computational studies that distinguishes them from other theoretical approaches to the study of vision is the notion that an effective way to analyze the computational strategies used by biological systems to perform complex visual tasks is to build

computer vision systems that use similar strategies. The synthesis of machine vision systems allows rigorous testing of whether some strategy that is hypothesized for the biological system really works at solving an important problem in vision. Building machine vision systems often uncovers difficult aspects of a problem, and possible solutions, that were not realized upon first consideration.

Experimental studies from psychology and the neurosciences provide critical insights into the particular computations that underlie visual processing in biological systems. A given problem typically can be solved in more than one way; there are different choices for how properties of the physical world can constrain its solution, and different algorithms, or procedures, that can be used to carry out the solution. Computer implementations of possible strategies can show how the overall behavior of the system depends on the choice of constraints and algorithms, and critical experimental tests can be devised to determine what choices are made in biological systems. Thus, a fruitful interaction between computational and experimental approaches can contribute to our understanding of biological vision.

Low level and high level vision

Biological vision begins with measurements of the amount of light reflected from surfaces in the environment onto the eye. The retinal image provided by the photoreceptors can be thought of as a large array of continuously changing numbers that represent light intensities. From this array of light measurements, the visual system does not achieve an understanding of what is in the scene in a single step. A tenet of computational studies is that vision proceeds in stages, with each stage producing increasingly more useful descriptions of the world. The process of vision can be viewed as the construction of a series of representations of visual information with explicit computation that transforms one representation into the next.

It is not yet known how biological systems represent visual information, but computational studies have suggested several representations that are useful in visual processing (see, for example, Marr, 1982; Ballard and Brown, 1982; Horn, 1986; Fischler and Firschein, 1987). Representations proposed for the early stages of vision first capture information that can be extracted simply and directly from the initial image, such as the location and description of significant intensity changes or edges in the image. Subsequent representations capture the local geometry or three-dimensional shape of visible surfaces in the scene, represented as the orientation or depth of surfaces at each location in the scene. Many familiar visual processes, such as the analysis of movement, binocular stereopsis, surface shading, texture, and color, contribute to the computation of these early visual representations. We refer to these early stages of vision as *low level* vision. A main goal of low level visual processes is to recover properties of the surrounding environment, and the representations that they deliver can be evaluated on the basis of their validity, that is, whether the results they deliver are correct and accurate.

Later representations for vision capture information necessary to solve complex tasks such as navigation through the environment, manipulation of objects, and recognition. The visual processes involved in accomplishing these tasks are often referred to as *high level* vision. Although there is no well–defined boundary between low and high level vision, the distinction is often a useful one. The main difference between low and high level visual processes is in the kind of knowledge they use. Low level vision relies on assumptions regarding the general physical properties of objects, such as continuity and rigidity. High level tasks, such as recognition, often use some knowledge we have acquired about specific objects, such as their shape and perhaps the transformations that they may undergo. Additional distinctions between low and high level vision, and the description of an intermediate domain between them, are discussed in section 3 of this chapter.

Scope of the chapter

The next section focuses on three problem areas in which there has been substantial computational work, as well as important interactions between computational work and experimental studies from psychology and the neurosciences. In particular, we consider the problems of edge detection, binocular stereo and the analysis of visual motion. Overall, research in computational vision spans a wide range of problems that also include the analysis of texture, color, and the recovery of 3–D shape from image contours and smooth shading. Most early work explored these problems in isolation, while recent work has emphasized the integration of multiple visual processes. Reviews of computational work in these areas can be found in a number of recent books (for example, Brady, 1981; Ballard and Brown, 1982; Marr, 1982; Beck, Hope and Rosenfeld, 1983; Ullman and Richards, 1984; Levine, 1985; Horn, 1986; Pentland, 1986; Brown, 1987; Fischler and Firschein, 1987; Richards and Ullman, 1987). Section 3 explores problems in intermediate and high level vision, and addresses two major problems. The first is the extraction of shape properties of figures and spatial relations among items in the scene. The second is visual object recognition.

2. LOW LEVEL VISION

2.1 Edge Detection

When the image is first captured by the retinal photoreceptors in the eye, it consists of some 120 million individual pointwise measurements of light intensity. A major problem that arises is how to transform this large and unstructured set of measurements into a more useful representation, that is, a representation that is more concise and more convenient for subsequent processing stages.

In the fields of computer vision and image processing, the dominant suggestion has been to begin the analysis of the incoming image by producing from the intensity array an *edge*

representation, using a process called *edge detection*. Edges in the image are locations where the light intensity changes significantly from one level to a different one. The main rationale for using an edge representation as the primary representation is that intensity edges are usually physically significant: they correspond to object boundaries and to discontinuities in surface properties, such as spatial orientation or reflectance. The important role of edges is also supported by the fact that for humans, a line sketch of an image often conveys most of the essential information, although from the point of view of the underlying intensity distributions, the line sketch and the image are radically different. Similarly, in experiments involving pattern recognition by animals (such as the rat, octopus, and goldfish), it has been noted that figures are usually treated as equivalent to a sketch of their outlines (Sutherland, 1968). Early perceptual studies by Cornsweet and others (see, for example, Cornsweet, 1970) also show that our perception of an image does not directly mimic the initial intensities registered by the eye, but is strongly influenced by the presence of sharp intensity changes, or edges. An example of an edge representation obtained from a natural image, using an edge detection method for computer vision systems proposed by Canny (1986) is shown in Figure 1.

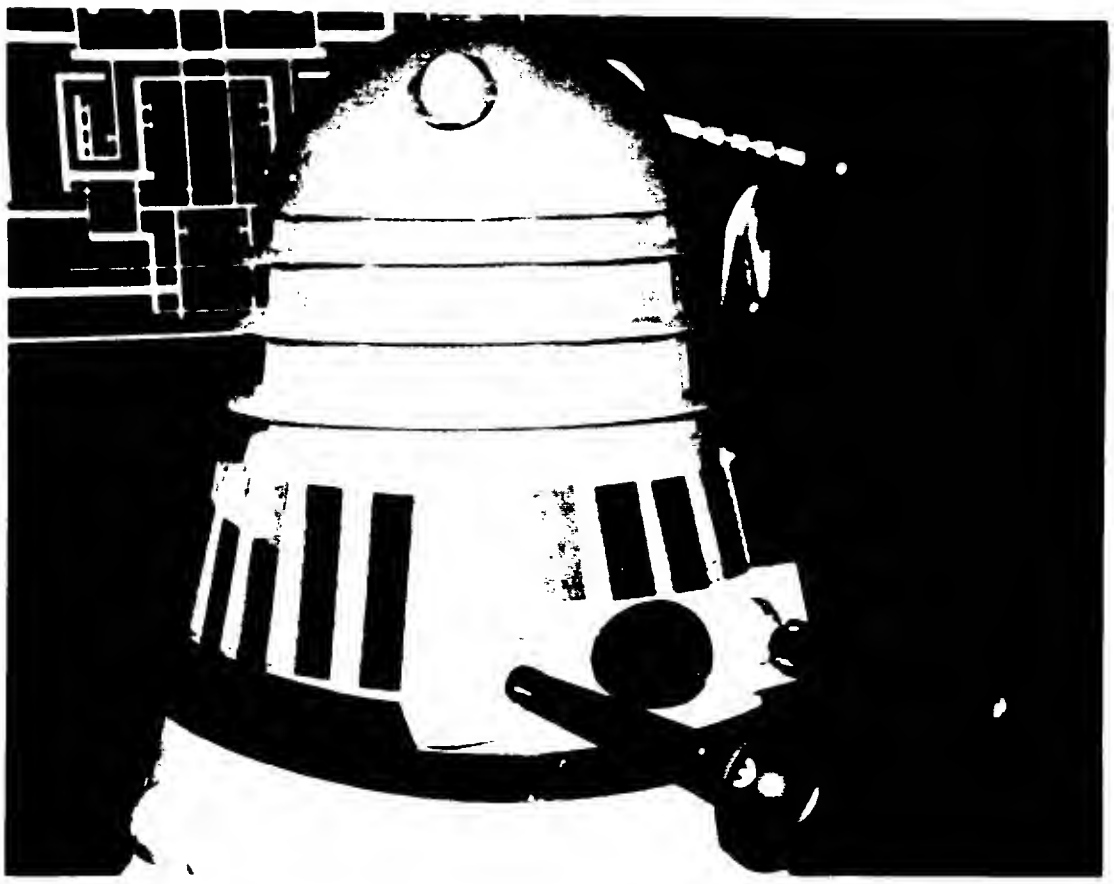
It should be mentioned that the operation of edge detection is not the only proposal that has been made concerning the initial representation of the incoming image. A popular alternative has been a local Fourier (or Gabor) decomposition of the image.[1] These alternative suggestions have not yet played as significant a role in computational vision, and will not be discussed further here.

In the discussion below we will consider primarily one family of edge detection schemes that are based on detecting maxima in the rate of change (defined as maxima, or peaks, in the first derivative or zero–crossings in the second) of the image intensities. The reasons for concentrating on this group are first, that such schemes have produced high quality results that have been used successfully in later stages of visual processing in computer vision systems, and second, interesting possible connections have been proposed between this scheme and edge detection by the human visual system. For reviews of various edge detection techniques, see, for example, Davis (1975), Fram and Deutsch (1975), Pratt (1978), Ballard and Brown (1982), Torre and Poggio (1986), Fischler and Firschein (1987), and Hildreth (1987).

---

[1] The description of the image using Fourier components is based on the mathematical notion that any periodic function can be expressed as the sum of more elementary functions (sines and cosines), each multiplied by an appropriate scaling factor (see, for example, Bracewell, 1978). The set of scaling factors (called the Fourier coefficients) can serve to characterize the function uniquely. The Fourier approach to early vision suggests that the incoming image is divided into regions, and for each region the visual system constructs a representation that resembles a Fourier decomposition.

a.



b.

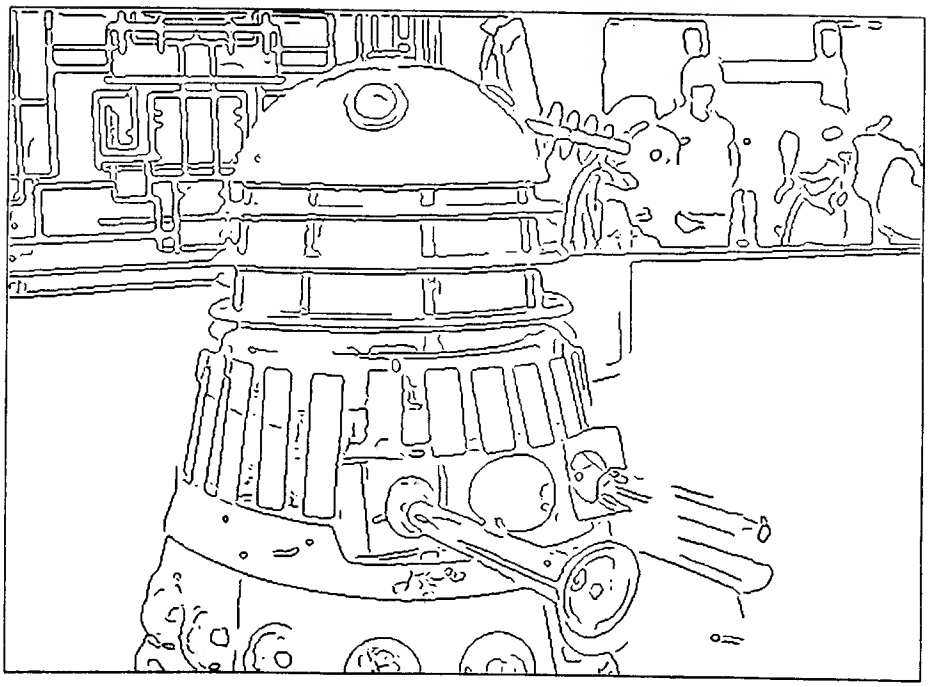


Figure 1. Edge detection. (a) Original image. (b) Edge representation of the same image, produced by the Canny edge detection algorithm (Canny, 1986).

Differentiation and smoothing

To produce an edge representation of the image, one needs a more precise definition of what is meant by "significant" intensity changes. One definition that has been proposed is to identify edges as locations in the image where the (directional) derivative of intensity reaches a local maximum in its absolute value. It can be shown that, under quite general conditions, when a signal passes through an optical system, step edges in the incoming intensity distribution (before the optics) will give rise to maxima in the first directional derivative of image intensity (after the optics). Locating derivative maxima therefore appears to be a reasonable step toward identifying sharp edges.

The differentiation of a signal, however, raises certain technical difficulties, because this operation enhances considerably the high frequency components of the signal. Because images often contain high frequency noise, the noise is enhanced by differentiation.[2] To overcome this problem, the differentiation of a signal is usually preceded by a smoothing operation that removes the noisy high frequency components. It has been shown that optimal results can be obtained by using a smoothing function that has a Gaussian, or close to a Gaussian shape (Shanmugam, Dickey and Green, 1979; Marr and Hildreth, 1980; Koenderink, 1984; Babaud *et al.*, 1986; Canny, 1986; Torre and Poggio, 1986). Thus, one approach to edge detection is to perform the following steps: (1) smooth the signal using a Gaussian filter, (2) differentiate the signal, and (3) locate maxima in its directional derivatives.[3]

Mathematically, these stages can be summarized (for a one–dimensional signal) as locating maxima in $\frac{d}{dx}(G * I)$, where G is a Gaussian function and $*$ denotes the filtering, or convolution operation. Interestingly, the same operation can be obtained by performing $(\frac{d}{dx}G) * I$. This means that the two successive operations, smoothing and differentiation, can be combined. The image is filtered through a new function, the first derivative of a Gaussian. Maxima in the first derivative of the smoothed image can be detected equivalently by finding locations where the second derivative of the smoothed image crosses zero. If we again use Gaussian smoothing, then mathematically, this second scheme corresponds to finding zero–crossings in the signal $(\frac{d^2}{dx^2}G) * I$. To summarize, edges can be found by (1) passing the image through a filter whose shape is the first or second derivative of a Gaussian, and (2) locating features such as maxima or zero–crossings in the result of this filtering stage. For the case of one dimension, the use of maxima and zero–crossings in the derivatives of Gaussian–filtered images is illustrated in Figure 2. The figure shows a one–dimensional intensity profile obtained by taking a slice through a natural image, the same

[2] Because the derivative of a sine function $\sin(\omega x)$ is $\omega \cos(\omega x)$, it follows that the effect of differentiation is to enhance periodic components with high spatial frequency (large ω).

[3] One can detect maxima in the absolute value of this derivative, or equivalently, maxima or minima (positive or negative *peaks*) in the derivative itself.

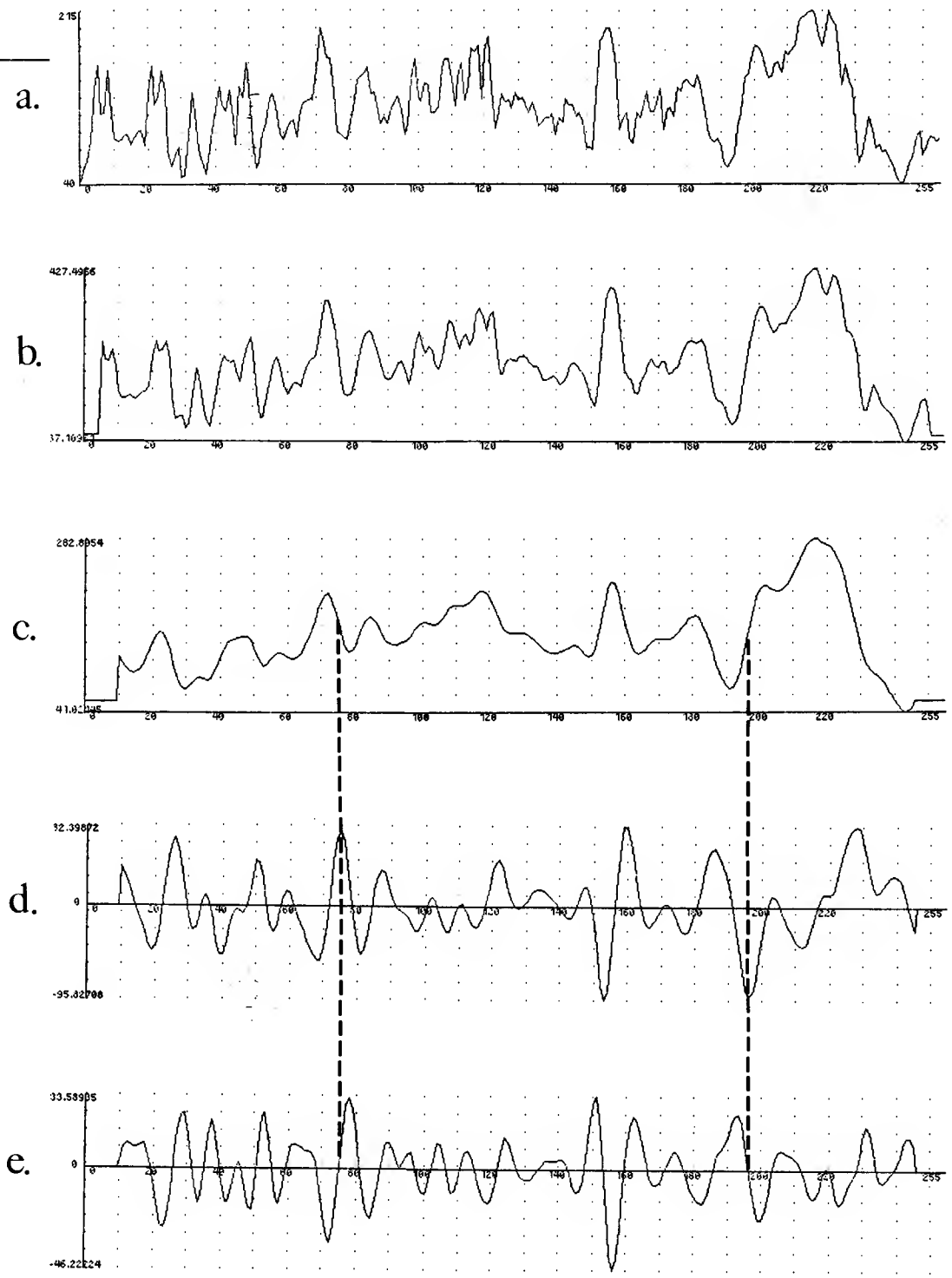


Figure 2. Detecting intensity changes. (a) One–dimensional intensity profile obtained by measuring intensities along a horizontal slice of a natural image. (b) The result of smoothing the profile in (a). (c) The result of additional smoothing of (a). (d) and (e) The first and second derivatives, respectively, of the smoothed profile shown in (c). The vertical dashed lines indicate peaks in the first derivative and zero–crossings in the second derivative that correspond to two significant intensity changes (Hildreth, 1987).

profile following Gaussian filtering, and how the edges can be identified by the maxima (or peaks) in the first derivative or zero–crossings in the second.

Peaks and zero–crossings in one and two dimensions

This basic scheme, presented above for a one–dimensional signal, has a number of possible extensions to the analysis of two–dimensional images. One alternative uses directional first or second derivative operators[4] for performing the two–dimensional differentiation (see, for example, Macleod, 1972; Davis, 1975; Persoon, 1976; Rosenfeld and Kak, 1976; Pratt, 1978; Haralick, 1980; Binford, 1981; Nevatia and Babu, 1980; Zucker, 1987). A second strategy is to evaluate, at each location in the image, the direction of the intensity gradient, which is the direction along which the intensity changes most rapidly. A one–dimensional analysis of intensity changes is then applied locally in this direction (for example, Canny, 1986). In a third alternative, the image is first passed through a single, non–directional filter, whose shape is defined by $\nabla^2 G$, the Laplacian of a two–dimensional Gaussian function (the Laplacian is the sum of the two directional derivatives in the horizontal and vertical directions), and then followed by the detection of the zero–crossings in the result of this filtering stage. The $\nabla^2 G$ operator is shown in Figure 3. It is interesting to note in this regard that under quite general conditions, the zero–crossings representation produced in this manner gives a complete representation of the image; that is, the original image can be reconstructed (up to a single scaling factor) from the zero–crossing contours alone (Curtis and Oppenheim, 1987).

Multiple scales

Intensity edges can occur in an image at a variety of scales. Some edges are sharp and close together, others are gradual and well separated. It proved difficult to capture all of the different types of significant intensity changes using only a single fixed operator. An approach that has evolved in computational vision in response to this problem is to analyze the image at a number of different scales (for example, Rosenfeld and Thurston, 1971; Marr and Hildreth, 1980; Mayhew and Frisby, 1981; Witkin, 1983; Canny, 1986). In a coarse scale edge representation, only a limited number of significant, relatively isolated edges would be revealed, whereas a fine scale (high resolution) analysis would yield a denser representation of the edges. In terms of the edge detection scheme outlined above, the resolution or scale of the analysis is determined by the size of the operator (such as the value σ of the space constant of the underlying Gaussian in the $\nabla^2 G$ operator). Multiple scale analysis is therefore achieved by the simultaneous use of a number of operators of different sizes, as shown in Figure 4.

[4]These operators compute the derivative of the smoothed image along a particular two–dimensional direction.

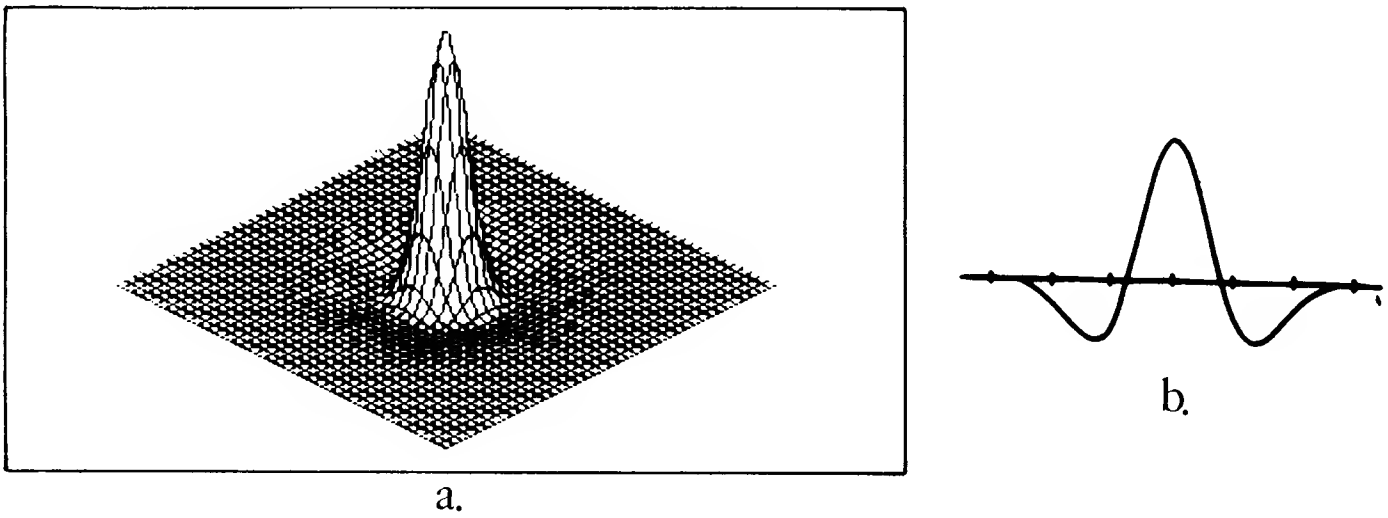
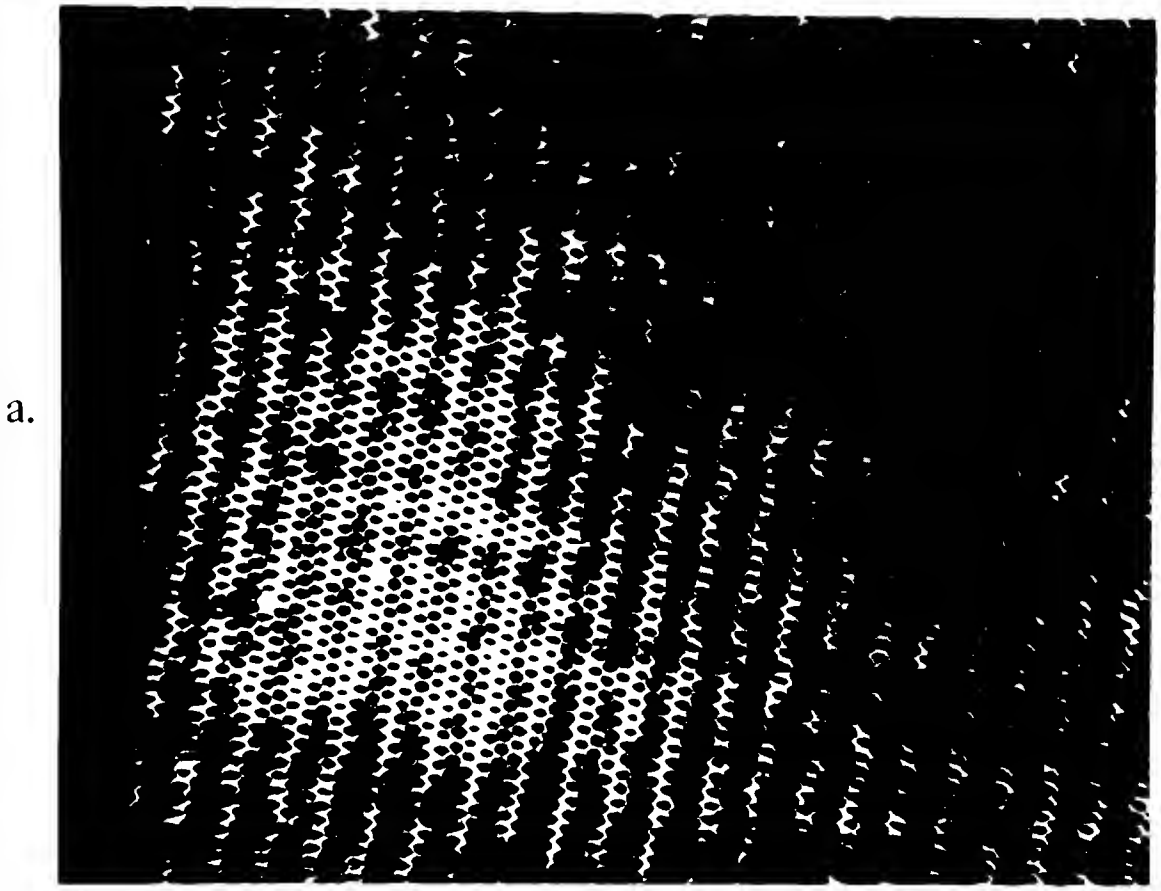


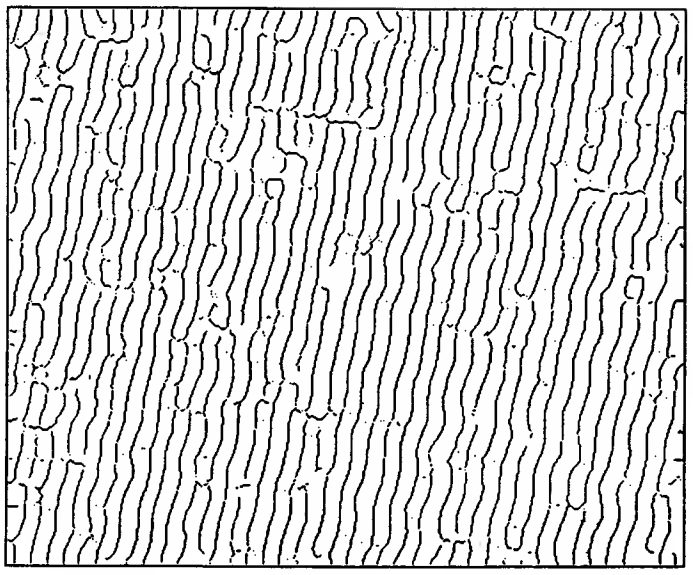
Figure 3. The $\nabla^2 G$ operator. (a) Three-dimensional plot of the shape of the two-dimensional operator. (b) A one-dimensional cross-section through the center of the operator.

This general idea of multiple scale analysis is in agreement with the large body of psychophysical data regarding multiple channels in the visual system. Experiments using techniques such as detection, adaptation, and discrimination, suggest that at each location in the visual field, a number of mechanisms analyze the image at a range of different scales (for example, Campbell and Robson, 1986; Cowan, 1977; Graham, 1977; Watson and Nachmias, 1977; Wilson and Bergen, 1979). These mechanisms appear to be sensitive to different ranges of spatial and temporal frequency.

Biological significance

One question regarding edge representations based on features such as peaks and zero-crossings is their biological relevance. It is known that the appropriate kind of filtering takes place at the level of the retina and the lateral geniculate nucleus. That is, at this level, the image can be thought of as being passed through filters, or receptive fields, whose shape can be closely approximated by the $\nabla^2 G$ shape presented above. It is not entirely clear, however, what processing takes place beyond this first stage, at the level of the primary visual cortex. It is, in fact, an instructive point that in spite of the large body of knowledge regarding the anatomy and physiology of the primary visual cortex, relatively little is known about the functions it performs. Many cortical studies are in general agreement with the notion that this area may be involved, among other functions, with the task of edge detection. In particular, the so-called cortical simple cells have often been described as edge and bar detectors. This view is not universally accepted, however (see, for example, DeValois,





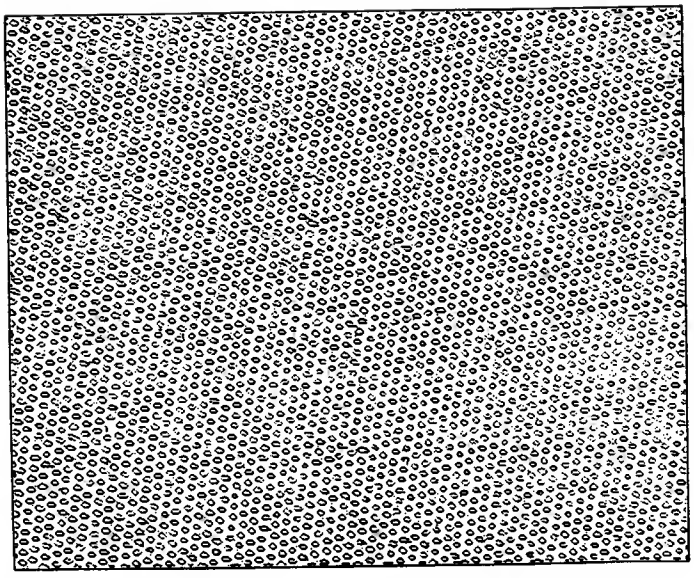


Figure 4. Multi–resolution image analysis. (a) The original image. (b) and (c) Edge representations obtained at two different scales (Canny, 1986).

Albrecht and Thorell, 1982), and the exact function of these and other cortical units remains unclear.

A possibility raised by the preceding discussion is that one of the functions of this cortical area may be to construct a number of zero–crossing maps at different scales, and then to combine them to construct an edge representation of the incoming image. This point of view regarding the possible function of the primary visual cortex gives rise to specific predictions that can be tested by physiological techniques. Some of these predictions have, in fact, been tested (Richter and Ullman, 1986), and the results appear generally to be compatible with a zero–crossing based view. Recent psychophysical studies (for example, Watt and Morgan, 1983, 1984) provide some support for the use of zero–crossings in human vision, but show that additional primitive features, such as stationary points in an approximation to the second derivative of the image intensities, are essential to account for the ability of the human visual system to detect and localize the positions of significant intensity changes in the image. The experiments performed so far are insufficient to determine conclusively the role of edges in early vision, but they serve to demonstrate that connections are starting to develop between computational theories on the one hand, and empirical biological studies of brain mechanisms on the other.

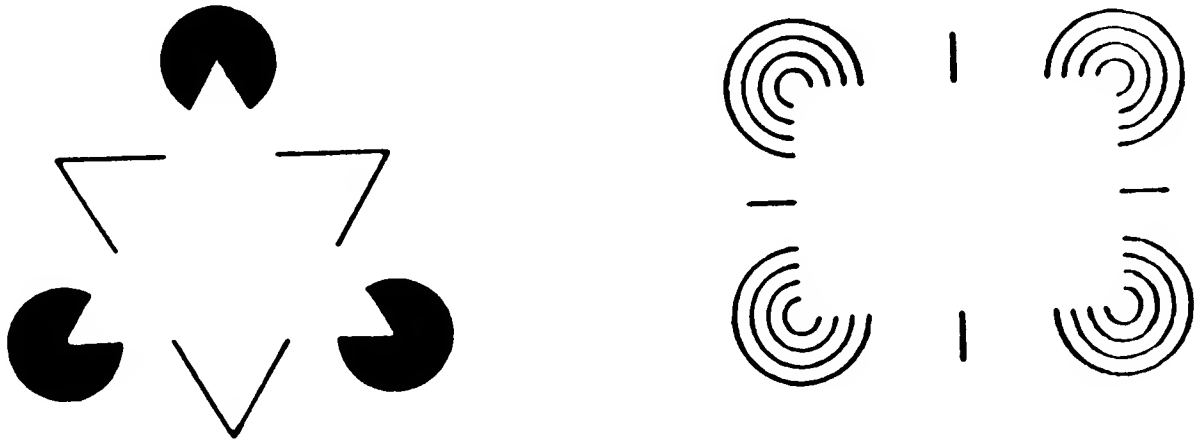


Figure 5. Subjective contours. Examples of boundaries that can be perceived, but which cannot be detected directly by an intensity–based edge detection scheme.

A final comment before leaving the topic of edge detection is that even if features such as peaks or zero–crossings in the result of smoothing and differentiating the image are being detected, it should be realized that they constitute only one stage in the analysis of boundaries. As mentioned above, edge representations at different scales must somehow be combined. In addition, there are many types of boundaries that humans are sensitive

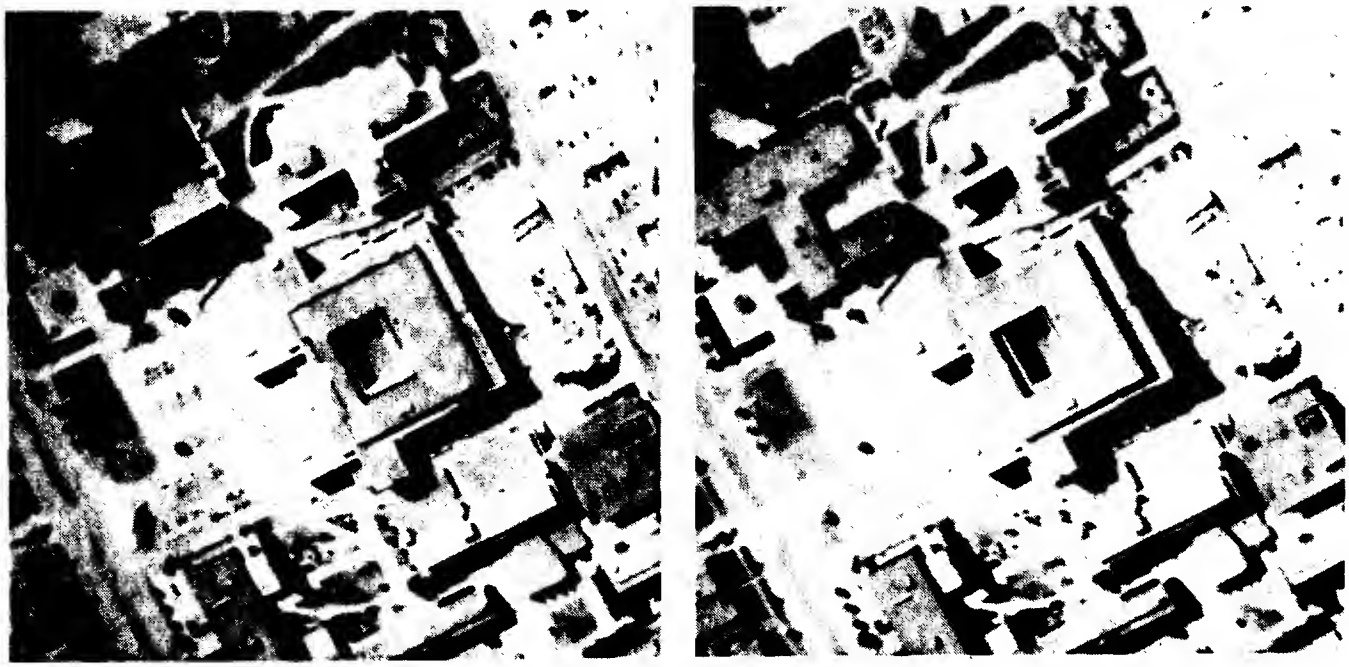
to, but which cannot be captured by the zero–crossing scheme or by other intensity–based edge detection techniques. Examples of such boundaries include texture edges, in which the two sides of the boundary differ in texture rather than in average intensity, and subjective contours, which the visual system fills in from different types of evidence when a physical discontinuity in the image is not present (see Figure 5). It is interesting to note that recently, cells have been found rather early in the visual system, in cortical area V2, that clearly respond to subjective contours (von der Heydt *et al.*, 1984).

2.2 Binocular Stereo

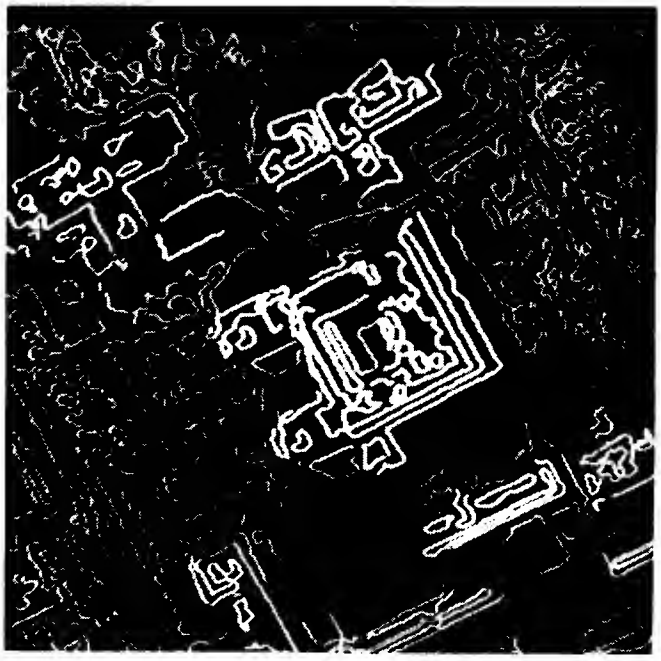
The left and right eyes obtain two slightly different views of the world, and we use the difference between these two views to recover the distances to surfaces in the environment. In particular, there is *disparity* between the positions of corresponding features in the two retinal images, which depends both on where the objects are located in space, and where the eyes are focused. Binocular stereo, the process that uses these disparities to recover 3–D information, is a critical component of human vision — it is considered our most accurate system for recovering the distances to surfaces (Westheimer and McKee, 1978, 1980; Poggio and Poggio, 1984).

Examining the process of stereo vision from a computational perspective, there are three main steps that need to be solved. The first is to extract elements from the images, whose positions will be compared in the two stereo views. The second is to match up these elements in the two views; that is, for each element in the left image, an element is located in the right image that corresponds to the same physical feature in the scene. Finally, through a geometric transformation, the disparities in position of features in the two images determine the distances to objects in space. It turns out that the second step, referred to as the *stereo correspondence problem*, is one of the most difficult aspects of the stereo process.

When we consider stereo views of a natural scene, such as those shown in Figure 6a, it might appear that the stereo correspondence problem ought to be straightforward. One could, for example, extract small patches from the left image, and look for the most similar patch of image intensities in the right image. This strategy is referred to as *grey–level correlation*, and it formed the basis for early stereo systems developed in computer vision (for review, see Barnard and Fischler, 1982), as well as early models of human stereo vision (for example, Sperling, 1970). Techniques based on matching the image intensities did not work well in practice, primarily for two reasons. First, the image intensities can change significantly between the two views — they undergo photometric and geometric distortions. An example of a photometric change between the two images is a bright highlight that appears in one image but not the other, because of the way the surface catches the light. An example of a geometric change is a rotation of one image relative to the other. Such rotations often occur when the angle of gaze deviates significantly from straight ahead. These distortions can lead to significant errors in stereo methods based on grey–level correlation.



a.



b.

Figure 6. Stereo analysis. (a) Aerial stereo photographs of a section of the campus at the University of British Columbia. (b) The disparities computed by the Marr–Poggio–Grimson stereo correspondence algorithm (Grimson, 1985), for the images shown in (a). Brightness encodes depth, with brighter contours being closer to the viewer.

There is a second problem, however, that is more severe — there is tremendous ambiguity in matching features between the two images. Given some patch in one image, there are often many patches in the other that are quite similar. This ambiguity is exacerbated by the wide range of depths contained in a typical scene, from which it follows that there will be a wide range of possible disparities between features in the two images. To solve the stereo correspondence problem, the matching process needs to be constrained, in order to derive a unique matching of features between the two stereo views.

The ambiguity problem is best illustrated with stereo images that consist of random dots, as shown in Figure 7 (Julesz, 1971). The left image is a random array of black and white dots. To construct the right image, patterns of dots from the left image are shifted by different amounts to the left or right in the right image, creating disparity in their positions in the two stereo views. In the case of Figure 7, a central patch of dots in the left image is shifted to the right in the right image. If the resulting patterns are observed in such a way that the left eye views only the left pattern and the right eye views the right, a perception of 3-D surfaces at different depths results from the relative placement of the dots. The ability to 'fuse' random-dot stereograms to form a perception of depth illustrates two important aspects of stereo vision. First, the features that are matched between the two images must be simple — stereo fusion does not require that we recognize objects in the image, and match entire objects between the two stereo views. Second, it illustrates the inherent ambiguity faced by the stereo correspondence process. In principle, any dot in the left image could match any dot in the right image of the same color.

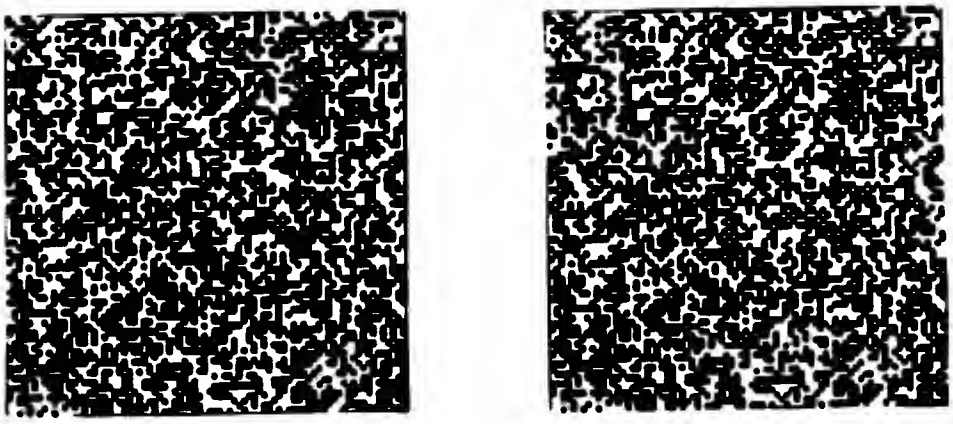


Figure 7. A random-dot stereogram with disparities that reveal two depth planes stacked on top of one another.

To solve the stereo correspondence problem uniquely, it is necessary to constrain the matching process in some way. Computational models of the stereo correspondence process use a variety of such constraints. First, an object in space occupies only a single location at one time, which implies that each feature in the left image matches only a single feature in the right image. This is referred to as the *uniqueness* constraint. Second, because of the nature of stereo projection, features located along a particular line in one image all match with features along a line in the other image. If the eyes are focused at a distance, then these corresponding lines are roughly horizontal and at the same vertical height in the two images. If, however, the eyes are focused at a nearby object, or the angle of gaze deviates significantly from being straight ahead, then these lines have different orientations in the two images. If the positions of the eyes in their orbit are known, then it is possible to determine where the corresponding lines lie in the two images. This facilitates stereo matching, because now, given a feature in one image, it is only necessary to search along a single line in the other image for a possible match. Virtually all models of stereo vision assume that the viewing geometry is known. This is referred to as the *epipolar* constraint.

Finally, all models of stereo correspondence assume some form of the constraint that the distance to surfaces in the scene tends to vary slowly across the image (for example, Marr and Poggio, 1976, 1979; Mayhew and Frisby, 1981; Baker, 1982; Barnard and Fischler, 1982; Medioni and Nevatia, 1985; Pollard, Mayhew and Frisby, 1985; Prazdny, 1986; Barnard, 1987; Hoff and Ahuja, 1987). Surfaces are usually smooth, so that the depths of nearby points on the surface are usually similar. Computational models have used this *continuity* constraint in different ways. Some try to match features in the two views in such a way that the variation in depth of the resulting surfaces is minimized (for example, Prazdny, 1986; Barnard, 1987). Others impose a constant limit on the rate of change of disparity across the image — in other words, they do not allow matches to be made that would imply a surface that is too steeply slanted in depth (for example, Medioni and Nevatia, 1985; Pollard, Mayhew and Frisby, 1985). This latter approach was motivated by psychophysical studies that suggest that in human stereo vision, there is a strict limit on the gradient of stereo disparity (for example, Tyler, 1975; Burt and Julesz, 1980). Finally, some models use a form of this constraint based on the idea that disparities vary slowly along connected edge contours in the image, referred to as *figural* continuity (Mayhew and Frisby, 1980, 1981; Baker and Binford, 1981; Baker, 1982; Medioni and Nevatia, 1985; Grimson, 1985). These latter models assume that features along continuous contours in the image are likely to lie along contiguous locations on a surface in space, and should therefore have similar or smoothly varying disparities. Mayhew and Frisby (1981) present psychophysical evidence that a figural continuity constraint is used in human stereo vision.

Simple strategies for computing stereo correspondence, such as the grey-level correlation technique mentioned earlier, work well on random-dot patterns and highly textured natural images, even when the underlying surfaces are quite complex. The need to provide a robust solution for a wide range of natural imagery has proven to be a difficult problem,

however, which appears to require a more complex strategy. In search of more successful algorithms, many computational models attempt to incorporate knowledge of human stereo vision. These models, in turn, provide insight into how these aspects of human vision are useful in solving the stereo correspondence problem.

There are at least three observations regarding human stereo vision that have been incorporated into computational models. First, as we noted in the previous section, there is much evidence that the human visual system initially extracts features in the image, such as sharp intensity changes or edges, and it has been suggested that these features may serve as the basic matching elements for stereo correspondence (for example, Barlow, Blakemore and Pettigrew, 1967; Marr and Poggio, 1979; Mayhew and Frisby, 1981). Second, perceptual demonstrations suggest that the human stereo system makes use of the different resolutions of analysis performed by the early stages of vision (for example, Felton, Richards and Smith, 1972; Julesz and Miller, 1975). Finally, it has been shown that eye movements are critical in fusing two stereo views. If the eyes are fixated at a particular distance, it is only possible to fuse features in the left and right views whose depths fall within a limited range around the fixation distance (known as Panum's fusional area; see, for example, Mitchell, 1966; Fender and Julesz, 1967). Objects outside of this range are seen as double. One method that can be used to fuse features over a wider range of depths is to perform vergence eye movements, fixating on different depth planes.

Many stereo models incorporate some or all of these factors. For example, most stereo systems match features between the two images, such as sharp intensity changes, edges, or other related features (for example, Marr and Poggio, 1979; Barnard and Thompson, 1980; Mayhew and Frisby, 1981; Grimson, 1981, 1985; Baker and Binford, 1981; Baker, 1982; Barnard and Fischler, 1982; Medioni and Nevatia, 1985; Pollard, Mayhew and Frisby, 1985; Prazdny, 1986; Hoff and Ahuja, 1987). Some models also assume that the corresponding features in the two eyes are similar to one another in their contrast, orientation in the image, sharpness, and so on, which can also narrow down possible matches (for example, Arnold and Binford, 1980; Baker and Binford, 1981; Baker, 1982; Medioni and Nevatia, 1985). Stereo correspondence models that match features obtained at multiple resolutions include those proposed by Marr and Poggio (1979), Moravec (1980), Grimson (1981, 1985), Mayhew and Frisby (1981), and Hoff and Ahuja (1987).

The analysis of appropriate matching features in computational models of stereo correspondence has motivated a number of psychophysical studies aimed at determining which features are used in human stereo vision. Mayhew and Frisby (1981) provide evidence that stereo matching features are extracted after the image has been filtered with a circularly symmetric operator (such as the circular difference–of–Gaussians function), and include the locations of peaks in this filtered image. Bulthoff and Mallot (1987) demonstrate that the human stereo system can derive a 3–D impression by matching smooth variations in shading, but that the sensation of depth is much weaker than that derived in the presence of sharp edge features in the stereo images.

The stereo correspondence model proposed by Marr and Poggio (1979) was presented as a possible model of human stereo vision, and has three main components. First, stereo matching takes place independently at different resolutions. Second, within each resolution, there is a limit to the amount of disparity over which features can be matched. At the coarsest scale, the algorithm searches over larger distances for matching features, whereas at the finest resolution, there is a more severe limit on the distance over which the image is searched for possible matches. Third, the matches established at the coarse resolution guide eye movements that shift the entire images right and left in a way that eventually allows the matching of features at all resolutions. Overall, matches at the coarse scale provide a rough idea of where objects are located in depth, and this rough depth map becomes successively refined as matches are established at finer scales. The system incorporates roles for multi–resolution analysis and eye movements, which are known to exist in human stereo vision. This algorithm (see also, Grimson, 1981, 1985) has been implemented in a number of computer vision systems, and is one of the most thoroughly tested of existing stereo systems. Figure 6b shows the results of this algorithm applied to the two aerial stereo photographs shown in Figure 6a. The brightness of edge contours represents depth, with brighter contours closer to the viewer. Computer models of this sort serve as a rigorous test for models of human stereo processing.

Other models of stereo correspondence have emerged over the last several years that also attempt to mimic aspects of the human stereo system (for example, Mayhew and Frisby, 1980, 1981; Pollard, Mayhew and Frisby, 1985). A key aspect of the early model proposed by Mayhew and Frisby (1981), for which they provide psychophysical evidence, is that the locations of image features at multiple spatial resolutions form a combined "signature" that is matched between the left and right views. This differs from the Marr–Poggio (1979) model, in which stereo matching takes place independently at different spatial scales. The model proposed by Pollard, Mayhew and Frisby (1985) (the PMF algorithm) combines the epipolar and uniqueness constraints with a constraint on the disparity gradient. That is, the rate of change of disparity from one image location to the next is restricted to lie within some specified bounds. The PMF algorithm has two stages. In the first, every potential match between a pair of features in the left and right images is assigned an initial "strength" that depends on the number of other potential matches within a small neighborhood whose disparities fall within the disparity gradient limit. The second stage is a relaxation procedure, in which matches that are unambiguous or which have maximum strengths are used to disambiguate other potential matches. The PMF algorithm was motivated by psychophysical studies that demonstrate the existence of a disparity gradient limit in human stereo vision (for example, Tyler, 1975; Burt and Julesz, 1980). Variations on the use of relaxation procedures for solving the stereo correspondence problem have also been considered by Marr and Poggio (1976), Barnard and Thompson (1981) and Prazdny (1985).

Given the importance of physical constraints in guiding the nature of computational solutions to the stereo correspondence problem, one would expect these constraints to be

reflected in the physiological mechanisms underlying binocular combination in biological systems. There has been considerable interest in understanding the neural circuitry that underlies stereo processing in the primate visual system (for review, see Poggio, 1984; Poggio and Poggio, 1984). In physiological studies by Poggio and his colleagues (Poggio and Fischer, 1977; Poggio and Talbot, 1981; Poggio, 1984), populations of cells were found in cortical areas V1 and V2 of the macaque monkey, which respond to the disparity in position of features appearing in the left and right eyes. Some cells respond best when objects appear in front of the point of fixation of the two eyes (NEAR cells), other cells respond best when objects appear behind the fixation point (FAR cells), and a third class responds to objects at depths right around the fixation point. Some of the latter class of cells are excited by disparities around zero (the tuned excitatory cells), while others are inhibited by these disparities (tuned inhibitory cells). It was shown recently that these cells respond to disparity in complex random-dot stereograms (Poggio, 1984), so their output effectively represents the result of solving a difficult stereo correspondence problem.

Several of the constraints used in computational models are reflected in the neural processes underlying stereopsis. For example, the left and right receptive fields of disparity sensitive cells are located at roughly the same vertical positions in the two eyes, so that these cells will be most active when the two stereo views are in approximate vertical register. (The human stereo system can tolerate large overall vertical disparities between the left and right images, but it appears that much of this disparity is compensated for through vertical eye movements that bring the two images into rough vertical register; only about $4'$–$7'$ of vertical disparity can be tolerated at a single eye position (Nielsen and Poggio, 1984).) The fact that corresponding features in the two eyes must be located at roughly the same vertical location is analogous to the assumption in computational models that the epipolar geometry is known.

The receptive fields of disparity sensitive cells are also located at roughly the same horizontal positions in the two eyes. For many of the NEAR and FAR cells, their disparity sensitivity for bar stimuli extends over a range of one degree or more on either side of zero disparity (Poggio and Fischer, 1977; Poggio and Talbot, 1981; Poggio, 1984). Others have a narrower depth response profile. This overall range is small compared to the actual range of disparity that typically exists for a natural scene. This mechanism effectively constrains the range of search between potential corresponding features in the two retinal images, for a given eye position. As noted earlier, vergence eye movements can be used to bring objects over a wider range of depths into roughly corresponding locations on the two eyes. Based on psychophysical evidence from stereo-anomalous observers, Richards (1971) proposed that human stereopsis combines the activity of three populations of neurons that are preferentially sensitive to crossed, near zero, and uncrossed disparities. The existence of the three classes of NEAR, FAR and TUNED disparity sensitive cells supports this hypothesis. The use of pools of disparity cells with limited disparity range is an explicit component of the Marr–Poggio (1979) stereo model. In addition, Marr and Poggio proposed that multiple spatial frequency

channels provide input to these disparity sensitive mechanisms, with coarser channels feeding mechanisms that are sensitive to larger ranges of disparity. This aspect of the model is supported by physiological studies of binocular mechanisms in the cat (Pettigrew *et al.*, 1968; Ferster, 1981), which show a strong correlation between receptive field size and width of disparity tuning curves.

Physiological evidence does not yet address the use of constraints such as smoothness and continuity of disparity information, or the use of a disparity gradient limit. These constraints presumably would be reflected in the nature of the interactions between neighboring disparity sensitive cells. Many of the computational models proposed for the correspondence process incorporate simple, local interactions between nearby disparity mechanisms, which could easily be implemented in parallel over the visual field (for example, Marr and Poggio, 1976; Barnard and Thompson, 1981; Prazdny, 1985; Pollard, Mayhew and Frisby, 1985; Poggio and Drumheller, 1986). Recent efforts have begun to address the details of the neural circuitry that might underlie the parallel computation of binocular disparity. Koch and Poggio (1987) proposed possible neural implementations of stereo disparity detectors, which yield specific predictions for their behavior that can be tested through physiological experiments.

Finally, we mention some other important aspects of stereo vision that have been addressed in computational and experimental studies, but which are not covered in this discussion. One problem is the detection of discontinuities in stereo disparity, which often occur at the locations of object boundaries. Such discontinuities violate the assumptions of smoothness and continuity of disparity, and may require specialized mechanisms to detect. A second problem is the "filling–in" of continuous surfaces between image locations where stereo disparities are initially computed. Human observers perceive continuous surfaces when observing sparse random–dot stereograms. Computational models that initially derive disparity measurements only at the locations of image features, such as edges, also raise the problem of interpolating a dense disparity map between sparse disparity information. A third problem is the transformation between stereo disparity and depth, which is likely to require knowledge about the current viewing geometry, such as the position of the eyes.

2.3 The Analysis of Visual Motion

The measurement and use of visual motion is one of the most fundamental abilities of biological vision systems, serving many essential functions. For example, a sudden movement in the scene might indicate an approaching predator or desirable prey. The rapid expansion of features in the visual field can signal an object about to collide with the observer. Discontinuities in motion often occur at object boundaries and can be used to carve up the scene into distinct objects. Motion signals provide input to centers controlling eye movements, allowing objects of interest to be tracked through the scene. Relative movement can be used to infer the 3–D structure and motion of object surfaces, and the movement

of the observer relative to the scene, allowing biological systems to navigate quickly and efficiently through the environment.

The pattern of movement in the changing image is not given to the visual system directly, but must be inferred from the changing intensities that reach the eye. The 3–D shape of object surfaces, the locations of object boundaries and the movement of the observer relative to the scene can in turn be inferred from the pattern of image motion. Typically, the overall analysis of motion is divided into these two stages: first, the measurement of movement in the changing two–dimensional (2–D) image, and second, the use of motion measurements, for example to recover the 3–D layout of the environment. The following two sections discuss computational studies that address these two aspects of motion analysis.

The measurement of motion

The measurement of movement may itself be divided into multiple stages and performed in different ways in biological systems. There is evidence that in the human visual system, motion may be measured by at least two different processes, termed *short–range* and *long–range* processes (Braddick, 1974, 1980). The short–range process analyzes continuous motion, or motion presented discretely, but with small spatial and temporal displacements from one moment to the next. The long–range process may then analyze motion over larger spatial and temporal displacements, as in apparent motion. There is evidence that these two processes interact at some stage (for example, Green, 1983), but initially they may be somewhat independent. This section addresses one fundamental aspect of the short–range measurement of motion, the solution of the *aperture problem.*

Consider the computation of a projected 2–D velocity field, which assigns to each feature in the image, a direction and speed of velocity. Figure 8a shows an example of the projected velocity field for a 3–D wireframe object that is rotating around a central vertical axis. In both biological and computer vision systems, the first mechanisms for measuring this motion examine only a limited region of the visual image, and as a consequence, often provide only partial information about the 2–D velocity field, due to the *aperture problem* (Wallach, 1976; Fennema and Thompson, 1979; Burt and Sperling, 1981; Horn and Schunck, 1981; Marr and Ullman, 1981; Adelson and Movshon, 1982). This problem arises when an oriented pattern in the image, such as an edge contour, extends beyond the region of the image analyzed by the initial motion detection mechanisms. The information provided by such mechanisms is illustrated in Figure 8b. In this case, an extended edge **E** moves across the image and its movement is observed through a window defined by the circular aperture **A**. Through this window, it is only possible to observe the movement of the edge in the direction perpendicular to its orientation. The component of motion along the orientation of the edge is invisible through this limited aperture. Thus it is not possible to distinguish between motions in the directions **b**, **c** and **d**. This problem is inherent in any motion detection operation that examines only a limited area of the image, and arises when an

oriented pattern extends beyond the region of the image analyzed by the initial motion detection mechanisms.

As a consequence of the aperture problem, the measurement of motion in the changing image requires two stages of analysis: the first measures components of motion in the direction perpendicular to image features; the second combines these components of motion to compute the full 2–D pattern of movement in the image. In Figure 8c, a circle undergoes pure translation to the right. The arrows along the contour represent the perpendicular components of velocity that can be measured directly from the changing image. These component measurements each provide some constraint on the possible motion of the circle, as illustrated in Figure 8d. The bold vector **v** represents the local perpendicular component of motion at a particular location in the image. The possible true motions at that location are given by the set of velocity vectors whose endpoint lies along the line l oriented perpendicular to the vector **v**. Examples of possible true velocities are indicated by the dotted vectors. The movement of image features such as corners or small spots can be measured directly. In general, however, the first measurements of movement provide only partial information about the true movement of features in the image, and must be combined to compute the full pattern of 2–D motion.

The measurement of movement is difficult because in theory, there are infinitely many patterns of motion that are consistent with a given changing image. For example, in Figure 8e, the contour **C** rotates, translates and deforms to yield the contour **C′** at some later time. The true motion of the point **p** is ambiguous. Additional constraint is required to identify a unique solution. In general, it may not be possible to recover the 2–D projection of the true 3–D field of motions of points in space, from the changing image intensities. Factors such as changing illumination, specularities and shadows can generate patterns of optical flow in the image that do not correspond to the real movement of surface features. The additional constraint used to measure image motion can yield at best a solution that is most plausible from a physical standpoint.

Many physical assumptions could provide the additional constraint needed to compute a unique pattern of image motion. One possibility is the assumption of pure translation. That is, it is assumed that velocity is constant over small areas of the image. This assumption has been used both in computer vision studies and in biological models of motion measurement (for example, Lappin and Bell, 1976; Pantle and Picciano, 1976; Fennema and Thompson, 1979; Anstis, 1980; Marr and Ullman, 1981; Thompson and Barnard, 1981; Adelson and Movshon, 1982). Methods that assume pure translation may be used to detect sudden movements or to track objects across the visual field. These tasks may require only a rough estimate of the overall translation of objects across the image. Tasks such as the recovery of 3–D structure from motion require a more detailed measurement of relative motion in the image, and require the use of a more general physical assumption.

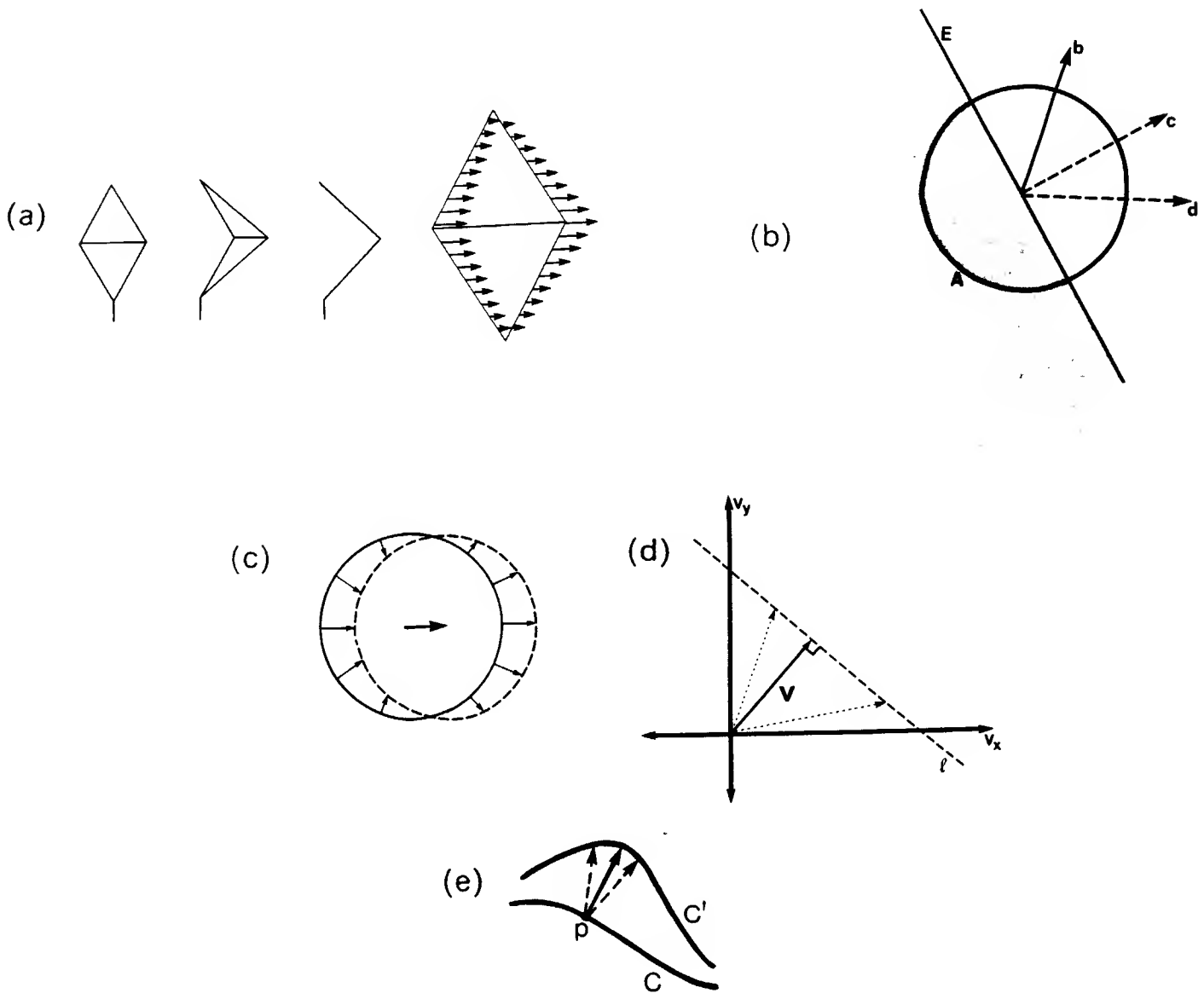


Figure 8. The aperture problem. (a) On the left are three views of a wireframe object undergoing rotation around a central vertical axis. On the right, the arrows along the contours of the object represent the instantaneous velocities of points along the object's contours at one position in the object's trajectory. (b) An operation that views the moving edge **E** through the local aperture **A** can compute only the component of motion **c** in the direction perpendicular to the orientation of the edge. (c) The circle undergoes pure translation to the right; the arrows represent the perpendicular components of velocity that can be measured from the changing image. (d) The curve **C** rotates, translates and deforms over time to yield the curve **C'**. The velocity of the point **p** is ambiguous. (e) The vector **v** represents the perpendicular component of velocity at some location in the image. The true velocity at that location must project to the line l perpendicular to **v**; examples are shown with dotted arrows.

Other computational studies have assumed that velocity varies smoothly across the image (Horn and Schunck, 1981; Hildreth, 1984; Nagel, 1984; Nagel and Enkelmann, 1984; Anandan and Weiss, 1985). The assumption rests on the principle that physical surfaces are generally smooth; that is, variations in the structure of a surface are usually small, compared with the distance of the surface from the viewer. When surfaces move, nearby points tend to move with similar velocities. There exist discontinuities in movement at object boundaries, but most of the image is the projection of relatively smooth surfaces. Thus, it is natural to assume that image velocities vary smoothly over most of the visual field. A unique pattern of movement can be obtained by computing a velocity field that is consistent with the changing image and has the least amount of variation possible.

The use of the smoothness constraint allows general motion to be analyzed. Surfaces can be rigid or nonrigid, undergoing any movement in space. This assumption also can be embodied in the motion measurement computation in a way that guarantees a unique solution, which is often physically correct (Hildreth, 1984; Ullman and Yuille, 1987). Finally, the velocity field of least variation can be computed straightforwardly, using standard computer algorithms (Horn and Schunck, 1981; Hildreth, 1984; Nagel and Enkelmann, 1984), as well as simple analog resistive networks that resemble the properties of neural networks (Poggio, Torre and Koch, 1985; Poggio and Koch, 1985). Perceptual studies reveal a similarity between the behavior of a motion measurement computation based on the smoothness assumption and the human perception of motion (Hildreth, 1984; Nakayama and Silverman, 1987a,b).

The aperture problem in motion measurement can be examined from a physiological perspective. Early movement detectors in biological systems have spatially limited receptive fields, and therefore face the aperture problem whenever an extended, oriented pattern moves across their receptive fields. Stimulated by a theoretical analysis of this problem, Movshon *et al.* (1985) sought and found direct physiological evidence for a two–stage motion measurement computation in the primate visual system. Two visual areas that include an abundance of motion–sensitive neurons are cortical areas V1 and MT. MT is the middle temporal area of extrastriate cortex, located in the posterior bank of the superior temporal sulcus (Van Essen and Maunsell, 1983). Movshon *et al.* (1985) explored the type of motion analysis taking place in area MT, using visual stimuli that consist of superimposed sinewave gratings, with different orientations and directions of motion. The results of these experiments indicate that the selectivity of neurons in area V1 for direction of movement is such that they could provide only the component of motion in the direction perpendicular to the orientation of image features. Area MT, however, contains a subpopulation of cells, referred to as *pattern* cells, that appear to respond to the 2–D direction of motion of a combined grating pattern, independent of its individual components. These neurons may serve to combine motion components to compute the real 2–D direction of velocity of a moving pattern. These experiments do not yet distinguish between the use of the simple assumption of pure translation, as suggested in the study (Movshon *et al.*, 1985), versus a more

general assumption such as smoothness. Stimulus patterns undergoing more complicated motions are required to make such a distinction. If the pattern cells in area MT embody the assumption of smoothness in their computation of motion, one would expect to find direct interaction between pattern cells that analyze nearby areas of the visual field.

The recovery of 3–D structure from motion

When an object moves in space, the motions of individual points on the object differ in a way that conveys information about its 3–D structure. Using 2–D shadow projections of 3–D wireframe objects, such as that shown in Figure 8a, Wallach and O'Connell (1953) showed that the human visual system can derive the correct 3–D structure of moving objects from their changing 2–D projection alone. Other perceptual studies also demonstrated this remarkable ability (for example, Braunstein, 1976; Johansson, 1975; Rogers and Graham, 1979; Ullman, 1979). Relative motion in the image is also created by movement of the observer relative to the environment, and can be used to infer observer motion from the changing image (for example, Gibson, 1950; Johansson, 1971; Lee, 1980; Regan, Kaufman and Lincoln, 1986).

Theoretically, the two problems of (1) recovering the 3–D structure and movement of objects in the environment and (2) recovering the 3–D motion of the observer from the changing image, are closely related. The main difficulty faced by both is that infinitely many combinations of 3–D structure and motion could give rise to any particular 2–D image. To resolve this inherent ambiguity, it is necessary to impose additional constraint that allows most 3–D interpretations to be ruled out, leaving one that is most plausible from a physical standpoint. Computational studies have used the *rigidity* assumption to derive a unique 3–D structure and motion; they assume that if it is possible to interpret the changing 2–D image as the projection of a rigid 3–D object in motion, then such an interpretation should be chosen (for example, Ullman, 1979, 1983; Clocksin, 1980; Prazdny, 1980, 1983; Longuet–Higgins and Prazdny, 1981; Tsai and Huang, 1981; Mitiche, 1986; Waxman and Ullman, 1985; Waxman and Wohn, 1987). When the rigidity assumption is used in this way, the recovery of structure from motion requires the computation of the rigid 3–D object that would project onto a given 2–D image. The rigidity assumption was suggested by perceptual studies that described a tendency for the human visual system to choose a rigid interpretation of moving elements (Wallach and O'Connell, 1953; Gibson and Gibson, 1957; Jansson and Johansson, 1973; Johansson, 1975, 1977).

Computational studies have shown that the rigidity assumption can be used to derive a unique 3–D structure from the changing 2–D image. Furthermore, this unique 3–D interpretation can be derived by integrating image information only over a limited extent in space and in time. For example, suppose that a rigid object in motion is projected onto the image plane by using orthographic projection. Three distinct views of four points on the moving object are sufficient to compute a unique rigid 3–D structure for the points

(Ullman, 1979). In general, if only two views of the moving points are considered or fewer points are observed, there are multiple rigid 3–D structures consistent with the changing 2–D projection. Theoretical results regarding the recovery of a unique 3–D structure under a variety of conditions are summarized in Ullman (1983) and Hildreth and Koch (1987). These theoretical results are important for the study of the recovery of structure from motion in biological vision systems, for two reasons. First, they show that by using the rigidity assumption, a unique structure can be recovered from motion information alone. It is not necessary to make further physical assumptions, in order to obtain a unique solution. Second, these results show that it is possible to recover 3–D structure by integrating image information over a small extent in space and in time. This second observation could bear on the neural mechanisms that compute structure from motion; in principle, they need only integrate motion information over a limited area of the visual field and a limited extent in time.

Computational studies also provide algorithms for deriving the structure of moving objects. Typically, measurements of the positions or velocities of image features give rise to a set of mathematical equations whose solution represents the desired 3–D structure. The algorithms generally derive this structure from motion information extracted over a limited area of the image and a limited extent in time. Testing of these algorithms reveals that although this strategy is possible in theory, it is not reliable in practice. A small amount of error in the image measurements can lead to very different (and often incorrect) 3–D structures. This behavior is due in part to the observation that over a small extent in space and time, very different objects can induce almost identical patterns of motion in the image (Ullman, 1983, 1984a).

This sensitivity to error inherent in algorithms that integrate motion information only over a small extent in space and time suggests that a robust scheme for deriving structure should use image information that is more extended in space or time or both. This conclusion is supported in recent computational studies (for example, Bruss and Horn, 1983; Lawton, 1983; Ullman, 1984a; Adiv, 1985; Bolles and Baker, 1985; Negahdaripour and Horn, 1985; Yasumoto and Medioni, 1985; Bharwani, Riseman and Hanson, 1986; Shariat and Price, 1986; Subbarao, 1986; Waxman and Wohn, 1987), which show that consideration of motion information that is more extended in space or time can lead to a stable recovery of structure. The extension in time can be achieved by considering a large number of discrete frames or by observing continuous motion over a significant temporal extent.

With regard to the human visual system, the dependence of perceived structure on the spatial and temporal extent of the viewed motion has not yet been studied systematically, but the following informal observations have been made. Regarding spatial extent, two or three points undergoing relative motion are sufficient to elicit a perception of 3–D structure (Borjesson and von Hofsten, 1973; Johansson, 1975), although theoretically the recovery of structure is less constrained for two points in motion, and perceptually the sensation of

structure is weaker. In addition, appropriate visual stimulation in a small region of the peripheral visual field is adequate to elicit a strong sense of self–motion (Johansson, 1971). Regarding the temporal extent of viewed motion, Johansson (1975) showed that a brief observation of patterns of moving lights generated by human figures moving in the dark (commonly referred to as biological motion displays) can lead to a perception of the 3–D motion and structure of the figures. Other perceptual studies indicate that the human visual system requires an extended time period to reach an accurate perception of 3–D structure (Wallach and O'Connell, 1953; White and Mueser, 1960; Doner, Lappin and Perfetto, 1984; Inada *et al.*, 1986). A brief observation of a moving pattern sometimes yields an impression of structure that is "flatter" than the true structure of the moving object. Thus, the human visual system is capable of deriving some sense of structure from motion information that is integrated over a small extent in space and time. An accurate perception of structure may, however, require a more extended viewing period.

Most methods compute a 3–D structure from motion only when the changing image can be interpreted as the projection of a rigid object in motion. They otherwise yield no interpretation of structure or yield a solution that is incorrect or unstable. Algorithms that are exceptions to this can interpret only restricted classes of nonrigid motions (for example, Bennett and Hoffman, 1985; Hoffman and Flinchbaugh, 1982; Koenderink and van Doorn, 1986; Subbarao, 1986). The human visual system, however, can derive some sense of structure for a wide range of nonrigid motions, including stretching, bending and more complex types of deformation (Johansson, 1975; Jansson and Johansson, 1973; Todd, 1982, 1984). Furthermore, displays of rigid objects in motion sometimes give rise to the perception of somewhat distorting objects (Wallach, Weisz and Adams, 1956; White and Mueser, 1960; Braunstein and Andersen, 1984; Hildreth, 1984; Adelson, 1985). These observations suggest that while the human visual system tends to choose rigid interpretations of a changing image, it probably does not use the rigidity assumption in the strict way that previous computational studies suggest.

Ullman (1984a) proposed a more flexible method for deriving structure from motion that interprets both rigid and nonrigid motion. Referred to as the *incremental rigidity scheme*, this algorithm uses the rigidity assumption in a different way from previous studies. It maintains an internal model of the structure of a moving object that consists of the estimated 3–D coordinates of points on the object. The model is continually updated as new positions of image features are considered. Initially, the object is assumed to be flat, if no other cues to 3–D structure are present. Otherwise, its initial structure may be determined by other cues available, from stereopsis, shading, texture, or perspective. As each new view of the moving object appears, the algorithm computes a new set of 3–D coordinates for points on the object that maximizes the rigidity in the transformation from the current model to the new positions. This is achieved by minimizimg the change in the 3–D distances between points in the model. Thus the algorithm interprets the changing 2–D image as the projection of a moving 3–D object that changes as little as possible from one moment

to the next. Through a process of repeatedly considering new views of objects in motion and updating the current model of their structure, the algorithm builds up and maintains a 3–D model of the objects. If objects deform over time, the 3–D model computed by the algorithm also changes over time. A parallel model based on Ullman's incremental rigidity scheme was recently proposed by Landy (1986).

Physiological studies have uncovered neurons in higher cortical areas that are sensitive to properties of the motion field that may be relevant to the recovery of the 3–D structure and motion of surfaces in the environment, or to the recovery of the motion of the observer relative to the scene. Many studies have revealed neurons sensitive to uniform expansion or contraction of the visual field, a property that is correlated either with translation of the observer forward or backward, or equivalently, motion of an object toward or away from the observer. Such neurons have been found, for example, in the posterior parietal cortex of the monkey (Motter and Mountcastle, 1981; Andersen, 1987). Other neurons have been found that are sensitive to global rotations in the visual field (see, for example, Andersen, 1987; Sakata *et al.*, 1985). All of these neurons have large receptive fields, so they probably lack the spatial sensitivity required to derive the detailed shape of an object surface from relative motion. A recent study by Siegel and Andersen (1986) shows that motion processing in area MT of primate visual cortex is critical to the recovery of structure from motion.

3. HIGH LEVEL VISION

The processes we have examined so far belong to the realm of low level vision. Their goal is primarily to recover properties of the surrounding environment. Such processes can be evaluated on the basis of their validity, that is, whether the results they deliver (such as depth, surface orientation, etc.) are correct and accurate. The low level processes were also characterized by a bottom–up and parallel mode of processing. "Bottom–up" means that the processing depends on the visual stimulus, and not on the task being performed. "Parallel" here means spatial parallelism; that is, the same operations are being performed across the entire visual field or large parts of it.

In high level vision, the descriptions produced by the earlier stages are used for tasks such as recognition, visually guided manipulation, locomotion, and navigation through the environment. At this stage the emphasis is often on usefulness rather than validity. There are, for example, many possible ways to describe the elements in a given image for the purpose of recognition, but some may be more useful than others. The processing here is goal directed (as opposed to bottom–up), and spatially focused (as opposed to spatially uniform).

A further distinction can be made between intermediate and high level vision. High level vision uses specific knowledge about objects in the world, such as a catalog of objects stored in long term memory, while intermediate vision does not. The terms "intermediate"

and "high level" vision do not mean to imply a sequential order. It is possible, for example, that recognition tasks may proceed on their own without relying on the prior application of intermediate level processes. In section 3.1 we discuss the area of intermediate vision, and in particular, the problem of extracting shape properties and spatial relations among objects in the scene, from visual information. In section 3.2 we discuss one of the major problems in high level vision — visual object recognition.

3.1 Intermediate Vision and Visual Routines

The extraction of shape properties and spatial relations

The use of visual information often requires the extraction of shape properties of contours and regions and the analysis of spatial relations among items in the image. This analysis plays a role in visual classification and recognition. An example is shown in Figure 9. The figure on the left is recognized effortlessly and immediately as representing a face, as opposed to the jumbled face figure on the right. In this example, the features comprising the face are highly schematic. The eye or nose by themselves, for example, are crude representations of a real eye or nose. It is primarily the spatial relations among the schematic items that make it a face profile. The eye is inside the head boundary, the nose is attached to the boundary, the eye is above the nose and below the eyebrow, and so on. Our ability to recognize objects easily from such a representation demonstrates that we can naturally and effortlessly extract in a single glance spatial relations such as inside, above or attached–to, and then use these relations in recognition tasks.

There is some evidence that the ability to analyze such configurations and to detect certain basic abstract shape properties and spatial relations already exists at a very early age. For example, some studies (Fantz, 1961) have suggested that infants as young as one to five weeks of age already make a distinction between "correct" and jumbled schematic faces. At the physiological level, neurons were found in the STS area of the macaque visual cortex that respond specifically to faces (Gross, Rocha–Miranda and Bender, 1972). According to some reports (Perret, Rolls and Caan, 1982), these cells often respond to schematic line–drawn faces, but not to scrambled face figures. These findings suggest the existence of mechanisms that respond not only to specific and realistic faces, but to any configuration that has the general overall shape and correct spatial relations among its parts.

The visual analysis of shape and spatial relations also plays an important role in planning actions and manipulating objects. During such activities, we often use vision to answer such questions as "can object A fit into the space between objects B and C?" as shown in Figure 10. These questions may not be posed explicitly, but information about such relations is obtained routinely in our daily activities. Problems of this type do not require recognition, because they do not depend on naming the objects or on past familiarity with them. They do, however, require the spatial analysis of shape and relations among objects.

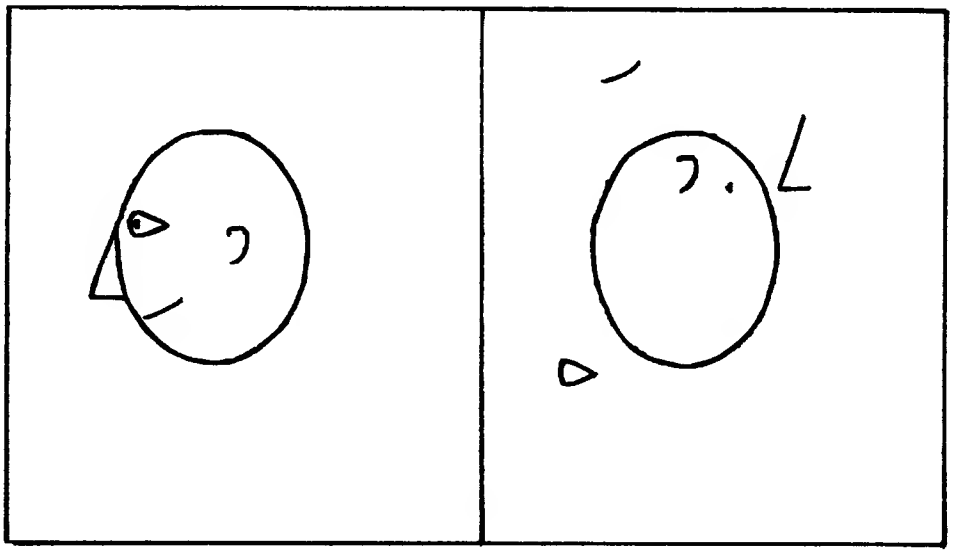


Figure 9. Recognizing faces. The figure on the left is perceived immediately as a face profile. The figure on the right contains similar parts in a different spatial arrangement.

In this sense, this analysis belongs more to the level termed above "intermediate vision" than to high level vision.

Properties of intermediate level vision: non-uniformity, open-endedness and task dependence

The computations involved in extracting shape and spatial relations are also clearly not part of the early visual processes, but must operate at a later, separate stage. As far as we know, the operations involved in the construction of the early visual representations (such as edge detection, the computation of depth, color, motion analysis, and so on), proceed uniformly and in parallel across the entire visual field or a large part of it. In contrast, it is not feasible to have at every location an "inside/outside detector," in the same way that orientation of contours is recovered by a dense array of specialized orientation detectors. More generally, it is not feasible to extract uniformly and in a bottom-up manner all the possibly relevant shape properties and spatial relations in the image.

The extraction of shape and spatial relations is also more open-ended than the earlier visual processes. That is, different tasks may require the extraction of different shape properties and relations. There does not seem to be a clear bound on the number of properties and relations that can be extracted; new ones can be learned as required by the task. As a result of these properties, the computations involved in intermediate level vision cannot be entirely bottom-up, but must be task dependent. That is, in viewing the same

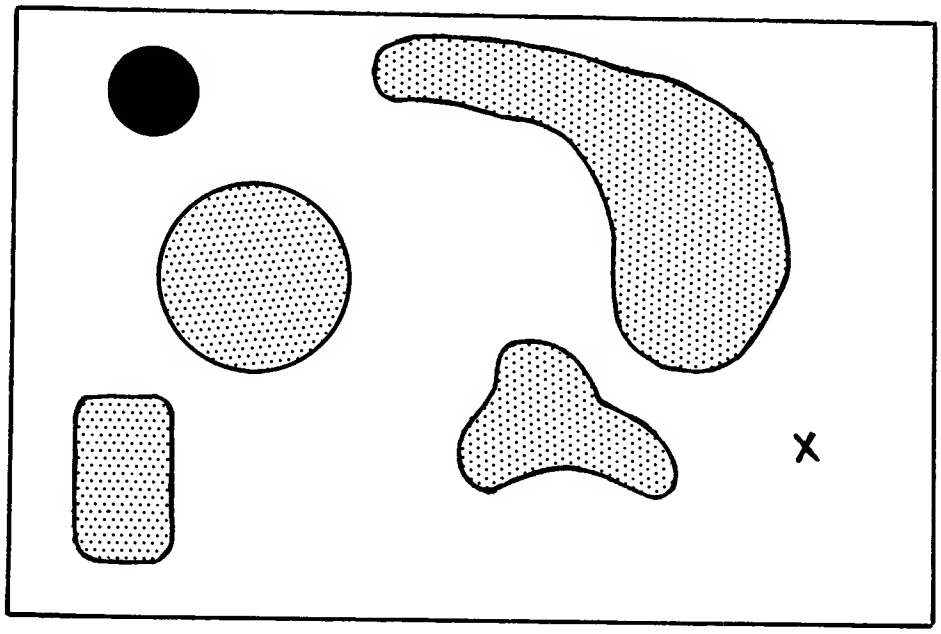


Figure 10. Spatial reasoning. It is possible to obtain, from visual information, answers to questions such as, can the black disk be moved to the location of the cross without colliding with any of the objects in the figure?

scene on two different occasions, the processes applied to the scene at this intermediate level of processing may be different, depending on the visual task being accomplished.

The complexity of extracting shape properties and spatial relations

The perception of many shape properties and relations appears to us subjectively to be immediate and effortless. This apparent immediateness is, however, deceiving; the computations required for extracting shape and spatial relations are often quite complicated. In fact, in many cases the efficient analysis of shape and spatial relations is still beyond the reach of current computer vision systems.

A simple illustrative example is the perception of inside/outside relations, that is, whether a given location is inside or outside a closed curve. When the curve is not too convoluted, this relation is easy to extract by "merely looking" at the image. In Figure 11, the location marked "x" is clearly inside the curve in Figure 11a, and outside the curve in Figure 11b. How do we reach this conclusion so effortlessly?

In computer vision, a popular method of computing inside/outside relations is the so–called "ray intersection" method. A ray is drawn from the location in question to the edge

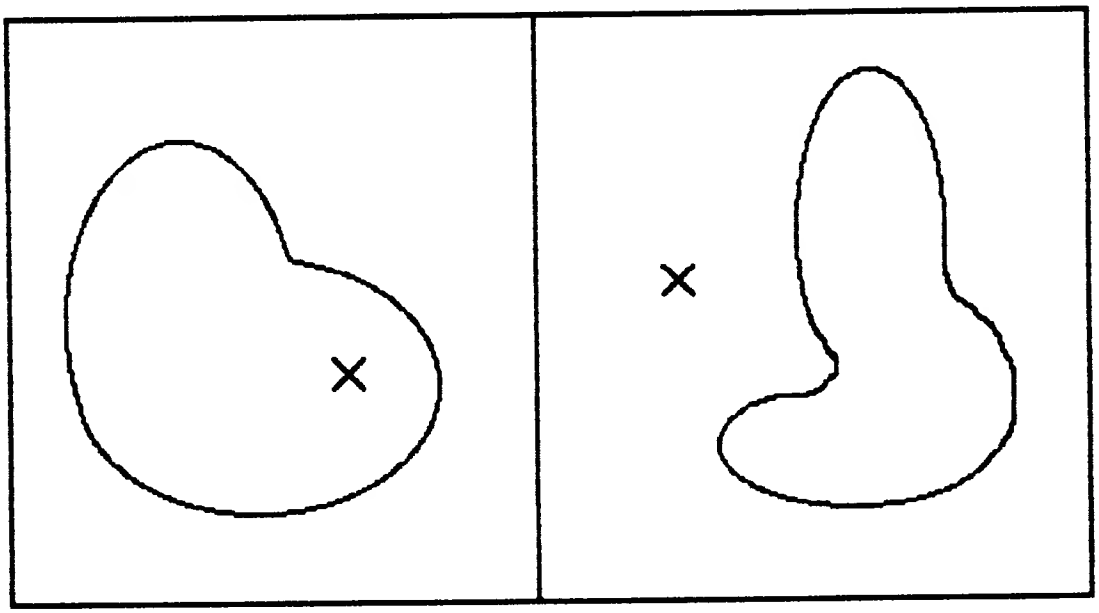


Figure 11. Judging inside/outside. The fact that the x lies inside the closed figure in (a) and outside in (b) is established effortlessly, by "merely looking" at the figure.

of the image, and the intersections made by the ray with the curve are counted. If the number is even, the point is outside the curve, otherwise it must be inside. This procedure is simple, but clearly insufficient. Consider, for instance, the examples in Figure 12. In Figure 12a, the number of ray intersections is even, but the point lies inside the curve. This is because the ray intersection method makes an implicit assumption that the curve in question is isolated in the visual field, an assumption that is violated in this example. In Figure 12b, the number of intersections is odd, therefore the ray intersection method would conclude that the point lies inside the curve. In this case, the error arises because the ray intersection method incorporates the assumption that the curve is closed. Figure 12c is a slight variation on the inside/outside problem. The task is to determine whether any of the points lie inside the closed curve. The ray intersection method would require applying the test to all of the individual dots. What our visual system seems to do instead is somehow to inspect the figure and verify that none of the dots lie in its interior. Finally, in Figure 12d, the task is to find an "innermost" location (that is, a point that lies inside all three curves). This task is visually straightforward, but it cannot be solved by any simple application of the ray intersection method.

The conclusion is that the method used by our own visual system is more powerful and flexible than the ray intersection method or any other method proposed so far in compu-

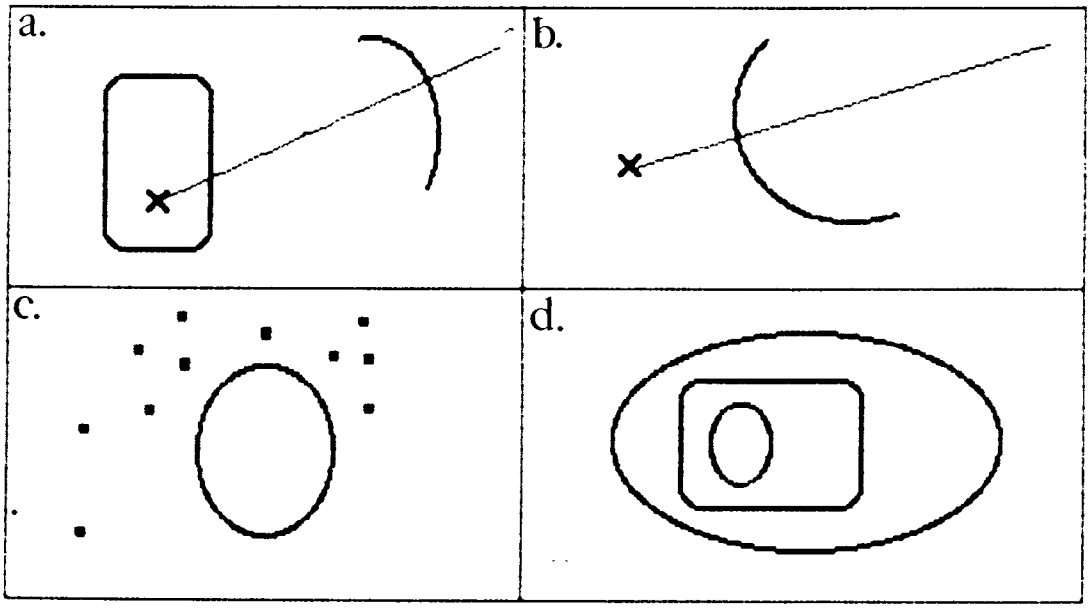


Figure 12. Deficiencies of the ray intersection method. (a) The number of intersections is even, but the point lies inside the curve. (b) The number is odd, but the point is not inside a closed region. (c) The task is to determine whether any of the points lies inside the figure. (d) The task is to find a location that is internal to all three curves.

tational vision. More generally, we see that the computation of seemingly simple spatial relations is surprisingly difficult. In addition, the problem is complicated by the facts that intermediate level vision requires the capacity to extract efficiently an essentially unbounded set of properties and relations, and that the computations performed must depend on the particular visual task at hand.

Visual routines

A possible approach to this problem is that the perception of shape properties and spatial relations is achieved by the application of *visual routines* to the early visual representation (Ullman, 1984b). These visual routines are sequences of some basic operations that are "wired into" the visual system. Routines for different properties and relations are composed from the same set of basic operations, using different combinations for different tasks. Using a fixed set of basic operations, the visual system can assemble different routines, and in this manner extract a large variety of different shape properties and relations. The basic operations themselves are carried out by specialized and highly efficient mechanisms. To understand the perception of spatial relations in general, an important step

would be to identify the set of basic operations. To explain how we perceive a particular relation such as "above" or "inside" would require, in this framework, a specification of the particular routine used for establishing the relation in terms of the underlying operations. In the following section, we describe briefly three basic operations that appear to play a useful role in visual routines: shifting and indexing, coloring, and boundary tracing.

Shifting and indexing

The application of visual routines implies a spatial and temporal structuring of the processing. A certain operation is applied at a given location in the visual field, then (perhaps depending on the results of this application) another operation is applied at a new location. "Shifting" is the general operation of moving the focus of processing to a required location. "Indexing" is a particular type of shifting, the shifting of the processing focus to a salient location in the visual field. There is a considerable body of experimental evidence related to the redirecting of the processing focus, the so–called "spotlight" of visual attention. This body of evidence is reviewed, for example, by Posner (1980), and Treisman and Gelade (1980).

A simple and elegant demonstration that appears to employ shifting and indexing comes from Treisman's experiments with the conjunction of elementary properties (see also Beck and Ambler, 1973). Her experiments have shown that certain odd–man–out targets can be detected in a field of distractors in a parallel manner: the time to detect the target does not depend significantly on the number of distractors in the background. For example, a target blue X can be detected in constant time in a field of brown T's and green X's (up to thirty background elements were used in these experiments) on the basis of its unique color. Similarly, a target T can be detected in a field of X's in constant time, presumably on the basis of orientation differences. In contrast, certain combinations (conjunctions) of elementary properties require sequential search. For example, the target may be a green T in a field of brown T's and green X's. The target in this case matches half the distractors in shape and half in color. The evidence suggests that in this case the processing becomes quite different: the visual system appears to scan the items sequentially in search of the required conjunction. The scan is performed to a large extent internally, rather than by eye movements.

The scan used in the conjunction search is a psychophysical example of the use of a shifting operation. In the parallel search situation, the odd–man–out element is immediately indexable, that is, the processing focus can shift immediately to it without dwelling on other elements. This is, therefore, an example of an indexing operation. Recently, phenomena related to shifting and selective visual attention have also been studied at the physiological level (see, for example, the study by Moran and Desimone (1985) regarding location–specific responses in area V4 of the alert macaque monkey).

Region coloring and boundary tracing

Coloring and tracing are two somewhat similar operations. "Coloring" is the operation of labeling a region with a unique label. Such labeling can be achieved by the bounded activation of a region in the following manner. Starting from a given point, the area around it in the internal representation is somehow activated. This activation spreads outward until a boundary is reached, but it is not allowed to cross the boundary. If the initial point lies inside a closed boundary, the entire region will be "colored," or labeled. This labeling can provide, for example, a basis for separating "inside" from "outside" in the tasks considered in Figures 11 and 12. These tasks also often involve an indexing operation: the x is indexed first, and this location serves as the starting point for the coloring operation. Additional stages would be required in all of these examples to complete the inside/outside computations; however, these additions will not be considered further here.

Boundary tracing is a similar operation applied to one–dimensional entities in the image, such as contours and boundaries. Starting from a given location on a contour, the contour is traced sequentially in a given direction and labeled. From a computational standpoint, such an operation could serve a useful role in the analysis of contours and boundaries, by separating them from nearby contours. For a general discussion of coloring and boundary tracing, see Ullman (1984b), and for a psychophysical exploration of boundary tracing, see Jolicoeur, Mackey and Ullman (1986).

We have considered above only a small sample of the problems in intermediate level vision. On the whole work in this area is only in its infancy, and many questions remain unanswered. The approach discussed above is only one of several possible alternatives and has so far been primarily motivated by computational considerations, and little is known at present about the processes by which the human visual system extracts abstract shape properties and spatial relations in the image.

3.2 Visual Object Recognition

The common view of object recognition is that it requires some sort of matching between the internal representations of an object stored in memory, and a similar representation generated from an object in the image. The process of object recognition is therefore different from processes of intermediate and low level vision in that it is more intimately related to the problems of memory organization, retrieval, expectations, and reasoning.

In this section we consider the problem of recognizing objects visually on the basis of their shape. Shape–based recognition is only part of the problem of visual object recognition, as other "paths" to recognition besides shape are possible. Objects are sometimes recognized more on the basis of their color or texture than on the basis of their shape. In other instances, visual recognition requires certain reasoning processes rather than the identification of a specific shape. We will also not consider here the problem often referred to

as the segmentation problem, which has to do with the delineation of the object of interest in an image. Finally, we will consider primarily the recognition of individual objects, rather than object classification.

Why is object recognition difficult? According to one approach (termed the "direct" approach (Abu-Mostafa and Psaltis, 1987)), the problem is primarily one of memory size and efficient parallel matching. To recognize objects, we have to store a large number of object views in memory. In attempting to recognize an object, we must compare the object currently in view with previously stored views of objects and select the one that resembles it most. Models of human associative memory (Willshaw, Buneman and Longuet-Higgins, 1969; Hopfield, 1982) have suggested that brain mechanisms can cope efficiently with both of these problems. In large scale neural networks, a large number of patterns can be stored, and a new pattern can be compared simultaneously with all previously stored patterns.

There are, however, at least two problems with the direct approach. First, the set of all possible views of all possible objects is likely to be prohibitively large, and at the same time redundant. To recognize, for instance, triangles of any shape, position, and orientation, it clearly is not necessary to store in memory a large number of representative shapes. Second, it should be noted that the comparison used in such schemes between the current image and the images stored in memory is a simple one, essentially a correlation of the viewed image with previously stored images. In contrast, we can recognize objects from novel views that may not be closely similar to any previously stored view, by this simple comparison measure.

The alternatives to the direct approach assume that we do not store and compare the input patterns per se. Instead, we produce from the image some sort of internal representation of the viewed object, and compare it with similar descriptions stored in memory. The hope is that some of the variations in the different appearances of an object will be handled by the description process. That is, an object seen under different viewing conditions, *e.g.* from different viewpoints, will still produce highly similar descriptions. Most of the schemes that have been proposed to date can be classified into three main approaches: the first based on invariant properties, the second on object decomposition into parts, and the third on a process of explicitly compensating for the transformation between the viewed object and the stored model.

The invariant properties approach

This approach to object recognition assumes that objects have certain invariant properties that are common to all of their views. Formally, a property of this type can be thought of as a function from object images to the domain of real numbers. For example, various applications have used the notion of a "compactness measure," defined as the ratio of the object's perimeter length to the square root of its apparent area. Round and compact objects will have a low score on this measure, while thin and elongated objects have a high

score. Furthermore, the measure is unaffected by rotation, translation, and scaling in the image plane. The idea is to define a number of such measures, and collectively they would serve to identify each object unambiguously. The overall recognition process is thus broken down into the extraction of a number of different properties, followed by a final decision stage based on these properties.

In a popular variant of this approach, a property is not expected to be entirely invariant for the different views of an object (or class of objects), but to be restricted to a small interval. Properties of different objects may have partially overlapping ranges, but it is hoped that by defining a number of different properties it will be possible to characterize each object uniquely. This leads naturally to the notion of "feature spaces" used extensively in pattern recognition. If n different properties are used, each object can be represented as a point in an n–dimensional space. Recognition and classification then become problems of clustering points in such spaces.

Another approach that belongs to this general class of invariant properties theories is Gibson's theory of high–order invariances (Gibson, 1950, 1979). Gibson suggested that invariant properties of objects may be discovered in "high order" invariances in the optic array. Such invariances were assumed to be based primarily on spatial and temporal gradients of texture density.

The invariant properties approach has been applied with some success to limited problems, such as the recognition of simple industrial parts under controlled viewing conditions. The approach does not generalize well, however, to more complicated domains. The weakness of the approach seems to be that it is difficult, and often impossible, to find relatively simple properties that are preserved across the different transformations that an object may undergo.

Object decomposition into parts

A second approach, which replaced the previous one as the main approach to object recognition, relies on the decomposition of objects into constituent parts. The approach appeals to the intuition that many objects seem to have natural parts. For example, the human body can be divided at some level into a head, torso, arms, and legs. In the decomposition approach, these parts are detected, and the recognition of the entire object proceeds on the basis of the constituent parts and their relationships.

There are a number of subclasses within this general approach, the most popular of which has been the "structural description" method. This approach assumes that it would be easier to capture object invariances after their parts have been identified. In particular, simple relations between parts will remain invariant across all object views. For example, in the character "T" one can define two main parts and certain relations (such as "join") that are common to all legal instances of this character.

An early example of a theory of this type applied to human vision is Milner's (1974) model of visual shape recognition. This theory deals primarily with 2–D shapes, and uses as basic parts primitive features such as edges and line segments. Marr and Nishihara (1978) proposed a structural description theory applied to 3–D objects. They used as primitive parts elongated volumes called "generalized cylinders." Recognition proceeds by constructing a description of the object in terms of these cylinders and their relationships. The description produced in this manner is "object centered," in the sense that the final description relies on the internal relations among parts, and not on their relations relative to the viewer. Additional recent theories of this class applied to human vision include Biederman's (1985) theory of recognition by components, and Hoffman and Richards' (1986) "codon" scheme for the description and recognition of image contours.

Although the use of object parts seems to have some clear merits, it also appears that the use of structural descriptions as proposed in most of these schemes has a number of limitations. One is that the decomposition into natural parts is often insufficient to characterize the object in question. In many cases precise shape, rather than the general arrangement of parts, appears to be important. Another problem that appears to limit the applicability of the approach is that many objects do not decompose naturally into the union of clearly distinct parts.

The alignment approach

The idea behind the alignment approach is to detect and explicitly compensate for the transformations between an object currently in view and its model stored in memory. This concept has been applied in the past only in simple domains, such as printed character recognition of a known font type (Neisser, 1967, chapter 3). In this application it is assumed that a character in view may differ from its ideal prototype stored in memory because it may rotate, translate, and change scale. The first step in the recognition process is therefore to compensate for these transformations. This compensation can be done prior to the identification of the letter in question. For example, displacement can be compensated for by computing the "center of mass" of the letter and then shifting the letter so that this point always coincides with a fixed location. Size and orientation can also be "normalized" during this first stage. Following the normalization, the remaining differences between the viewed letter and its stored model are small, and a straightforward comparison should suffice to identify the letter correctly.

This general approach has been revived recently and extended in an attempt to apply it to the recognition of complex 3–D objects. The general idea remains the same. The viewed object differs from its stored model, because it has undergone a certain transformation such as a change in position and 3–D rotation (in the case of rigid objects). In an alignment approach, the first stage is to detect this transformation prior to the identification of the object. The transformation can then be "undone"; this stage is called the *alignment* stage.

Following the alignment, the stored model and viewed object are in close agreement, and the correct model is therefore easier to determine.

Compared to translation, rotation and scaling in the image, the transformations that the image of a 3–D object can undergo are more complicated to determine and compensate for. A number of recent schemes (for example, Lowe, 1986; Ullman, 1986; Huttenlocher and Ullman, 1987) have been proposed for extending the alignment method to deal with natural objects. The approach appears to offer some advantages, but it has been demonstrated so far only in limited domains.

It must be concluded that in object recognition, which is one of the most fundamental aspects of human vision, theories (as well as experimental work) still have a long way to go. Some combination of an alignment process with part decomposition may offer a promising starting point, but these techniques must be extended and generalized considerably before they will be able to cope successfully with common objects that are recognized efficiently by the human visual system.

Acknowledgements: We thank Eric Grimson for comments on a draft of this manuscript.

REFERENCES

Abu–Mostafa, Y. S., Psaltis, D. 1987. Optical neural computing. *Scientific American*, 256(3):66–73.

Adelson, E. H. 1985. Rigid objects that appear highly non–rigid. *Invest. Ophthalmol. Vision Sci. Suppl.* 26:56.

Adelson, E. H., Movshon, J. A. 1982. Phenomenal coherence of moving visual patterns. *Nature* 300:523–525.

Adiv, G. 1985. Determining three–dimensional motion and structure from optical flow generated by several moving objects. *IEEE Trans. Pattern Anal. Machine Intell.* PAMI–7:384–401.

Anandan, P., Weiss, R. 1985. Introducing a smoothness constraint in a matching approach for the computation of optical flow fields. *Proc. IEEE Workshop on Computer Vision: Representation and Control*, Bellaire, MI, October, pp. 186–194.

Andersen, R. A. 1987. The anatomy and physiology of the inferior parietal lobule. In *Development of Spatial Relations*, ed. U. Belugi, Chicago: Univ. Chicago Press, *in press.*

Anstis, S. M. 1980. The perception of apparent motion. *Phil. Trans. R. Soc. London Ser. B* 290:153–168.

Arnold, R. D., Binford, T. O. 1980. Geometric constraints in stereo vision. *SPIE J.* 238:281–292.

Babaud, J., Witkin, A. P., Baudin, M., Duda, R. O. 1986. Uniqueness of the Gaussian kernel for scale–space filtering. *IEEE Trans. Patt. Anal. Machine Intell.* PAMI–8:26–33.

Baker, H. H. 1982. Depth from edge– and intensity–based stereo. PhD thesis, Univ. Ill., Urbana, IL.

Baker, H. H., Binford, T. O. 1981. Depth from edge– and intensity–based stereo. *Proc. 7th Intern. Joint Conf. on Artif. Intell.*, Vancouver, B.C., 631–636.

Ballard, D. H., Brown, C. M. 1982. *Computer Vision.* Englewood Cliffs, NJ: Prentice–Hall.

Barnard, S. T. 1987. A stochastic approach to stereo vision. In: *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms.* Fischler, M. A., Firschein, O., editors. Morgan Kaufmann: Los Altos, CA, 21–25.

Barnard, S. T., Fischler, M. A. 1982. Computational stereo. *Comput. Surveys* 14:553–572.

Barnard, S. T., Thompson, W. T. 1980. Disparity analysis of images. *IEEE Trans. Patt. Anal. Machine Intell.* PAMI–2, 333–340.

Barlow, Blakemore, C., Pettigrew, J. D. 1967. The neural mechanism of binocular depth discrimination. *J. Physiol.* 193:327–342.

Beck, J., Ambler, B. 1973. The effects of concentrated and distributed attention on peripheral acuity. *Percept. Psychophys.* 14(2):225–230.

Beck, J., Hope, B., Rosenfeld, A., editors. 1983. *Human and Machine Vision.* New York: Academic.

Bennett, B. M., Hoffman, D. D. 1985. The computation of structure from fixed axis motion: nonrigid structures. *Biol. Cybern.* 51:293–300.

Bharwani, S., Riseman, E., Hanson, A. 1986. Refinement of environmental depth maps over multiple frames. In: *Proc. IEEE Workshop on Motion: Representation and Analysis.* IEEE Computer Society: New York, 73–80.

Biederman, I. 1985. Human image understanding: Recent research and a theory. *Comp. Vis. Graph. Image Proc.* 32:29–73.

Binford, T. O. 1981. Inferring surfaces from images. *Artif. Intell.* 17:205–244.

Bolles, R. C., Baker, H. H. 1985. Epipolar–plane image analysis: a technique for analyzing motion sequences. In: *Proc. Third IEEE Workshop on Computer Vision: Representation and Control.* IEEE Computer Society: New York, 168–178.

Borjesson, E., von Hofsten, C. 1973. Visual perception of motion in depth: application of a vector model to three–dot motion patterns. *Percept. Psychophys.* 13:169–179.

Bracewell, R. N. 1978. *The Fourier Transform and its Applications.* New York: McGraw–Hill Book Co.

Braddick, O. J. 1974. A short–range process in apparent motion. *Vision Res.* 14:519–527.

Braddick, O. J. 1980. Low–level and high–level processes in apparent motion. *Phil. Trans. R. Soc. London Ser. B* 290:137–151.

Brady, J. M., editor. 1981. *Computer Vision.* Amsterdam: North–Holland.

Brady, J. M., Rosenfeld, A., editors. 1987. *Proceedings of the First International Conference on Computer Vision.* IEEE Computer Society: Washington.

Braunstein, M. L. 1976. *Depth Perception Through Motion.* New York: Academic Press.

Braunstein, M. L., Andersen, G. J. 1984. A counterexample to the rigidity assumption in the visual perception of structure from motion. *Perception* 13:213–217.

Brown, C., editor. 1987. *Advances in Computer Vision.* New Jersey: Erlbaum *(in press).*

Bruss, A., Horn, B. K. P. 1983. Passive navigation. *Comput. Vision Graph. Image Proc.* 21:3–20.

Bulthoff, H. H., Mallot, H. P. 1987. Interaction of different modules in depth perception. In: *Proc. First International Conference on Computer Vision.* Brady, J. M., Rosenfeld, A., editors. IEEE Computer Society: Washington, 295–306.

Burt, P., Julesz, B. 1980. A disparity gradient limit for binocular fusion. *Science* 208:615–617.

Burt, P., Sperling, G. 1981. Time, distance, and feature trade–offs in visual apparent motion. *Psych. Rev.* 88:171–195.

Campbell, F. W., Robson, J. G. 1968. Application of Fourier analysis to the visibility of gratings. *J. Physiol. London* 197:551–556.

Canny, J. 1986. A computational approach to edge detection. *IEEE Trans. Patt. Anal. Machine Intell.* PAMI–8:679–698.

Clocksin, W. F. 1980. Perception of surface slant and edge labels from optical flow: a computational approach. *Perception* 9:253–269.

Cornsweet, T. N. 1970. *Visual Perception.* New York: Academic Press.

Cowan, J. D. 1977. Some remarks on channel bandwidth for visual contrast detection. *Neurosci. Res. Prog. Bull.* 15:492–517.

Curtis, S., Oppenheim, A. 1987. Reconstruction of multidimensional signals from zero crossings. *J. Opt. Soc. Am.* 4(1):221–231.

Davis, L.S. 1975. A survey of edge detection techniques. *Comp. Graph. Image Proc.* 4:248–270.

De Valois, R., Albrech, D. G., Thorell, L. G. 1982. Spatial frequency selectivity of cells in macaque visual cortex. *Vis. Res.* 22:545–559.

Doner, J., Lappin, J. S., Perfetto, G. 1984. Detection of three–dimensional structure in moving optical patterns. *J. Exp. Psychol.: Human Percept. Perform.* 10:1–11.

Fantz, R. L. 1961. The origin of form perception. *Scientific American* 204(5):66–72.

Felton, B., Richards, W., Smith, A. 1972. Disparity processing of spatial frequencies in man. *J. Physiol.* 225:319–362.

Fender, D., Julesz, B. 1967. Extension of Panum's fusional area in binocularly stabilized vision. *J. Opt. Soc. Am.* 57:819–830.

Fennema, C. L., Thompson, W. B. 1979. Velocity determination in scenes containing several moving objects. *Comput. Graph. Image Proc.* 9:301–315.

Ferster, D. 1981. A comparison of binocular depth mechanisms in areas 17 and 18 of the cat visual cortex. *J. Physiol.* 311:623–655.

Fischler, M. A., Firschein, O., editors. 1987. *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms.* Los Altos, CA: Morgan Kaufman.

Fram, J.R., Deutsch, E.S. 1975. On the quantitative evaluation of edge detection schemes and their comparison with human performance. *IEEE Trans. Computers* C–24(6):616–628.

Gibson, J. J. 1950. *The Perception of the Visual World.* Boston: Houghton Mifflin.

Gibson, J. J. 1979. *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.

Gibson, J. J., Gibson, E. J. 1957. Continuous perceptive transformations and the perception of rigid motion. *J. Exp. Psychol.* 54:129–138.

Graham, N. 1977. Visual detection of aperiodic spatial stimuli by probability summation among narrow band channels. *Vis. Res.* 17:637–652.

Green, M. 1983. Inhibition and facilitation of apparent motion by real motion. *Vis. Res.* 23:861–865.

Grimson, W. E. L. 1981. *From images to surfaces: A computational study of the human early visual system.* Cambridge: MIT Press.

Grimson, W. E. L. 1985. Computational experiments with a feature based stereo algorithm. *IEEE Trans. Patt. Anal. Mach. Intell.* PAMI–7:17–34.

Gross, C. G., Rocha–Miranda, C. E., Bender, D. B. 1972. Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophysiol.* 35:96–111.

Haralick, R. M. 1980. Edge and region analysis for digital image data. *Comput. Graph. Image Proc.* 12:60–73.

Hildreth, E. C. 1984. *The measurement of visual motion.* Cambridge: MIT Press.

Hildreth, E. C. 1987. Edge detection. In: *Encyclopedia of Artificial Intelligence*, S. Shapiro, ed., New York: John Wiley, 257–267.

Hildreth, E. C., Koch, C. 1987. The analysis of visual motion: From computational theory to neuronal mechanisms. *Ann. Rev. Neurosci.* 10:477–533.

Hoff, W., Ahuja, N. 1987. Extracting surfaces from stereo images: An integrated approach. In: *Proc. First International Conference on Computer Vision.* Brady, J. M., Rosenfeld, A., editors. IEEE Computer Society: Washington, 284–294.

Hoffman, D. D., Flinchbaugh, B. E. 1982. The interpretation of biological motion. *Biol. Cybern.* 42:195–204.

Hoffman, D., Richards, W. 1986. Parts of Recognition. In: *From Pixels to Predicates*, ed. A.P. Pentland, Norwood NJ: Ablex.

Hopfield, J. J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci.* USA 79:2554–2558.

Horn, B. K. H. 1986. *Robot Vision.* Cambridge: MIT Press and McGraw–Hill.

Horn, B. K. P., Schunck, B. G. 1981. Determining optical flow. *Artif. Intell.* 17:185–203.

Hubel, D. H., Wiesel, T. N. 1962. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *J. Physiology, London*, 160:106–154.

Hubel, D. H., Wiesel, T. N. 1968. Receptive fields and functional architecture of monkey striate cortex. *J. Physiology,London*, 195:215–243.

Huttenlocher, D. P., Ullman, S. 1987. Object recognition using alignment. *Proc. Int. Conf. Comp. Vis.*, London, June, 102–111.

Inada, V. K., Hildreth, E. C., Grzywacz, N. M., Adelson, E. H. 1986. The perceptual buildup of three–dimensional structure from motion. *Invest. Ophthal. Visual Sci. Suppl.* 27:142.

Jansson, G., Johansson, G. 1973. Visual perception of bending motion. *Perception* 2:321–326.

Johansson, G. 1971. Studies on visual perception of locomotion. *Perception* 6:365–376.

Johansson, G. 1975. Visual motion perception. *Sci. Am.* 232:76–88.

Johansson, G. 1977. Spatial constancy and motion in visual perception. In *Stability and Constancy in Visual Perception*, ed. W. Epstein, New York: John Wiley.

Jolicoeur, P., Ullman, S., Mackay, M. 1986. Curve tracing: A possible basic operation in the perception of spatial relations. *Memory and Cognition.* 14(2):129–140.

Julesz, B. 1971. *Foundations of Cyclopean Perception.* Chicago: Univ. Chicago Press.

Julesz, B., Miller, J. 1975. Independent spatial–frequency–tuned channels in binocular fusion and rivalry. *Perception* 4:125–143.

Koch, C., Poggio, T. 1987. Biophysics of Computational Systems: Neurons, Synapses and membranes. In: *New Insights into Synaptic Function*, G. M. Edelman, W. E. Gall and W. M. Cowan, editors, Neurosciences Research Foundation and John Wiley and Sons, *in press.*

Koenderink, J. J. 1984. The structure of images. *Biol. Cybern.* 50:363–370.

Koenderink, J. J., van Doorn, A. J. 1986. Depth and shape from differential perspective in the presence of bending deformations. *J. Opt. Soc. Am. A* 3:242–249.

Landy, M. S. 1986. A parallel model of the kinetic depth effect using local computations. Mathematical Studies in Perception and Cognition Rep. No. 86–2, Department of Psychology, New York University, New York.

Lappin, J. S., Bell, H. H. 1976. The detection of coherence in moving random dot patterns. *Vision Res.* 16:161–168.

Lawton, D. T. 1983. Processing translational motion sequences. *Comput. Vision Graph. Image Proc.* 22:116–144.

Lee, D. N. 1980. The optic flow field: The foundation of vision. *Philos. Trans. R. Soc. London B* 290:169–179.

Levine, M. D. 1985. *Vision in Man and Machine.* New York: McGraw–Hill.

Longuet–Higgins, H. C., Prazdny, K. 1981. The interpretation of moving retinal images. *Proc. R. Soc. London Ser. B* 208:385–397.

Lowe, D. G. 1986. Three–dimensional object recognition from single two–dimensional images. *Robotics Research Technical Report 202*, Courant Institute of Math. Sciences, New York University.

Macleod, I. D. G. 1972. Comments on techniques for edge detection. *Proc. IEEE* 60:344.

Marr, D. 1982. *Vision.* San Francisco: Freeman.

Marr, D., Hildreth, E. C. 1980. Theory of edge detection. *Proc. R. Soc. London Ser. B* 207:187–217.

Marr, D. and Nishihara, H. K. 1978. Representation and recognition of the spatial organization of three–dimensional shapes. *Proc. Roy. Soc. B.* 200:269–291.

Marr, D., Poggio, T. 1976. Cooperative computation of stereo disparity. *Science* 194:283–287.

Marr, D., Poggio, T. 1979. A computational theory of human stereo vision. *Proc. R. Soc. London Ser. B* 204:301–328.

Marr, D., Ullman, S. 1981. Directional selectivity and its use in early visual processing. *Proc. R. Soc. London Ser. B* 211:151–180.

Mayhew, J. E. W., Frisby, J. P. 1980. The computation of binocular edges. *Perception* 9:69–86.

Mayhew, J. E. W., Frisby, J. P. 1981. Psychophysical and computational studies towards a theory of human stereopsis. *Artif. Intell.* 17:349–385.

Medioni, G. G., Nevatia, R. 1985. Segment-based stereo matching. *Comp. Vis. Graph. Image Proc.* 31:2–18.

Milner, P. M. 1974. A model for visual shape recognition. *Psychol. Rev.* 81(6):521–535.

Mitchell, D. E. 1966. Retinal disparity and diplopia. *Vision Res.* 6:441–451.

Mitiche, A. 1986. On kineopsis and computation of structure and motion. *IEEE Trans. Pattern Anal. Machine Intell.* PAMI–8:109–112.

Moran, J., Desimore, R. 1985. Selective attention gates visual processing in the extrastriate cortex. *Science.* 229:782–784.

Moravec, H. P. 1980. Obstacle avoidance and navigation in the real world by seeing a robot rover. *Stanford Artif. Intell. Lab. Memo 340.*

Motter, B. C., Mountcastle, V. B. 1981. The functional properties of the light-sensitive neurons in the posterior parietal cortex studied in waking monkeys: foveal spacing and opponent vector organization. *J. Neurosci.* 1:3–26.

Movshon, J. A., Adelson, E. H., Gizzi, M. S., Newsome, W. T. 1985. The analysis of moving visual patterns. In *Pattern Recognition Mechanisms*, ed. C. Chagas, R. Gattas, C. G. Gross. Rome: Vatican Press.

Nagel, H.–H. 1984. Recent advances in image sequence analysis. *Proc. Premier Colloque Image — Traitement, Synthese, Technologie et Applications*, Biarritz, France, May, pp. 545–558.

Nagel, H.–H., Enkelmann, W. 1984. Towards the estimation of displacement vector fields by "oriented smoothness" constraints. *Proc. 7th Int. Conf. on Pattern Recognition*, Montreal, Canada, July, pp. 6–8.

Nakayama, K., Silverman, G. H. 1987a. The aperture problem I: Perception of non-rigidity and motion direction in translating sinusoidal lines. *Vis. Res.*, in press.

Nakayama, K., Silverman, G. H. 1987a. The aperture problem II: Spatial integration of velocity information along contours. *Vis. Res.*, in press.

Negahdaripour, S., Horn, B. K. P. 1985. Direct passive navigation. *MIT Artif. Intell. Memo* 821.

Neisser, U. 1967. *Cognitive Psychology.* New York: Appleton–Century–Crofts.

Nevatia, R., Babu, R. 1980. Linear feature extraction and description. *Comp. Graph. Image Proc.* 13:257–269.

Nielsen, K. R. K., Poggio, T. 1984. Vertical image registration in stereopsis. *Vision Res.* 24:1133–1140.

Pantle, A. J., Picciano, L. 1976. A multistable display: evidence for two separate motion systems in human vision. *Science* 193:500–502.

Pentland, A. P., editor. 1986. *From Pixels to Predicates: Recent Advances in Computational and Robotic Vision.* Norwood, NJ: Ablex.

Perret, D. J., Rolls, E. T., Caan, J. 1982. Visual neurons responsive to faces in the monkey temporal cortex. *Exp. Brain Res.* 47:329–342.

Persoon, E. 1976. A new edge detection algorithm and its applications. *Comp. Graph. Image Proc.* 5:425–446.

Pettigrew, J. D., Nikara, T., Bishop, P. O. 1968. Binocular interaction on single units in cat striate cortex: Simultaneous stimulation by single moving slits with receptive fields in correspondence. *Exp. Brain Res.* 6:391–410.

Poggio, G. F. 1984. Processing of stereoscopic information in primate visual cortex. In: *Dynamic Aspects of Neocortical Function*, ed. G. Edelman, W. M. Cowan, W. E. Gall. New York: John Wiley.

Poggio, G. F., Fischer, B. 1977. Binocular interaction and depth sensitivity of striate and prestriate cortical neurons of the behaving rhesus monkey. *J. Neurophysiol.* 40:1392–1405.

Poggio, G. F., Talbot, W. H. 1981. Mechanisms of static and dynamic stereopsis in foveal cortex of the rhesus monkey. *J. Physiol.* 315:469–492.

Poggio, G. F., Poggio, T. 1984. The analysis of stereopsis. *Ann. Rev. Neurosci.* 7:379–412.

Poggio, T., Drumheller, M. 1986. On parallel stereo. *Proc. IEEE Intl. Conf. on Robotics and Automation*, 1986.

Poggio, T., Koch, C. 1985. Ill–Posed Problems in Early Vision: from Computational Theory to Analog Networks. *Proc. R. Soc. London B* 226:303–323.

Poggio, T., Torre, V., Koch, C. 1985. Computational Vision and Regularization Theory. *Nature*, 317: 314–319.

Pollard, S. B., Mayhew, J. E. W., Frisby, J. P. 1985. PMF: A stereo correspondence algorithm using disparity gradient limit. *Perception* 14:449–470.

Posner, M. I. 1980. Orienting of attention. *Quart. J. Exp. Psychol.* 32:3–25.

Pratt, W. 1978. *Digital Image Processing.* New York: John Wiley and Sons.

Prazdny, K. 1980. Egomotion and relative depth map from optical flow. *Biol. Cybern.* 36:87–102.

Prazdny, K. 1983. On the information in optical flows. *Comput. Vision Graph. Image Proc.* 22:239–259.

Prazdny, K. 1986. Detection of binocular disparities. *Biol. Cybern.*

Regan, D. M., Kaufman, L., Lincoln, J. 1986. Motion in depth and visual acceleration. In: *Handbook of Perception and Human Performance*, K. R. Boff, editor. Vol. 2, Chpt. 19, New York: Wiley.

Richards, W. 1971. Anomalous stereoscopic depth perception. *J. Opt. Soc. Am.* 61:410–414.

Richards, W., Ullman, S., editors. 1987. *Image Understanding 1985-86.* Ablex: Norwood, NJ.

Richter, J., Ullman, S. 1986. Nonlinearities in cortical simple cells and the possible detection of zero crossings. *Biol. Cyber.* 53:195–202.

Rogers, B. J., Graham, M. 1979. Motion parallax as an independent cue for depth perception. *Perception* 8:125–134.

Rosenfeld, A., Kak, A. 1976. *Digital Picture Processing.* New York: Academic Press.

Rosenfeld, A., Thurston, M. 1971. Edge and curve detection for visual scene analysis. *IEEE Trans. Comp.* C–20:562–569.

Sakata, H., Shibutani, H., Kawano, K., Harrington, T. L. 1985. Neural mechanisms of space vision in the parietal association cortex of the monkey. *Vision Res.* 25:453–464.

Schiller, P. H., Finlay, B. L., Volman, S. E. 1976. Quantitative studies of single cell properties in monkey striate cortex.(I) Spatio temporal organization of receptive fields. *J. Neurophysiol.*, 39:1288–1319.

Shanmugam, K. S., Dickey, F. M., Green, J. A. 1979. An optimal frequency domain filter for edge detection in digital pictures. *IEEE Trans. Patt. Anal. Machine Intell.* PAMI–1:37–49.

Shariat, H., Price, K. 1986. How to use more than two frames to estimate motion. In: *Proc. IEEE Workshop on Motion: Representation and Analysis.* IEEE Computer Society: New York, 119–124.

Siegel, R. M., Andersen, R. A. 1986. Motion perceptual deficits following ibotenic acid lesions of the middle temporal area (MT) in the behaving Rhesus monkey. *Neuroscience Abstr.* 12:324.8.

Sperling, G. 1970. Binocular vision: A physical and a neural theory. *Am. J. Psychol.* 83:461–534.

Subbarao, M. 1986. Interpretation of image motion fields: A spatiotemporal approach. In: *Proc. IEEE Workshop on Motion: Representation and Analysis.* IEEE Computer Society: New York, 157–165.

Sutherland, N. S. 1968. Outline of a theory of visual pattern recognition in animal and man. *Proc. Roy. Soc. London B* 171:297–317.

Thompson, W. B., Barnard, S. T. 1981. Lower–level estimation and interpretation of visual motion. *IEEE Computer*, August, pp. 20–28.

Todd, J. T. 1982. Visual information about rigid and nonrigid motion: A geometric analysis. *J. Exp. Psychol.* 8:238–252.

Todd, J. T. 1984. The perception of three–dimensional structure from rigid and nonrigid motion. *Percept. Psychophys.* 36:97–103.

Torre, V., Poggio, T. 1986. On edge detection. *IEEE Trans. Patt. Anal. Machine Intell.* PAMI–8: 147–163.

Treisman, A., Gelade, G. 1980. A feature integration theory of attention. *Cog. Psychol.* 12:97–136.

Tsai, R. Y., Huang, T. S. 1981. Uniqueness and estimation of three–dimensional motion parameters of rigid objects with curved surfaces. *Univ. Illinois Urbana–Champaign, Coordinated Science Laboratory Report* R–921.

Tyler, C. W. 1975. Spatial limitations of human stereoscopic vision. *SPIE J.* 120:36–42.

Ullman, S. 1979. *The Interpretation of Visual Motion.* Cambridge: MIT Press.

Ullman, S. 1983. Computational studies in the interpretation of structure and motion: summary and extension. In:*Human and Machine Vision*, ed. J. Beck, B. Hope, A. Rosenfeld. New York: Academic, 459–480.

Ullman, S. 1984a. Maximizing rigidity: the incremental recovery of 3-D structure from rigid and rubbery motion. *Perception* 13:255–274.

Ullman, S. 1984b. Visual routines. *Cognition.* 18:97–159.

Ullman, S. 1986. An approach to object recognition: Aligning pictorial descriptions. *MIT Artif. Intell. Lab. Memo.* 931.

Ullman, S., Richards, W., editors. 1984. *Image Understanding 1984.* Ablex: Norwood, NJ.

Ullman, S., Yuille, A. L. 1987. The smoothest velocity field. *MIT Artif. Intell. Memo.*

Van Essen, D. C., Maunsell, J. H. R. 1983. Hierarchical organization and functional streams in the visual cortex. *Trends Neurosci.*, 6:370–375.

von der Heydt, R., Peterhans, E., Baumgartner, G. 1984. Illusory contours and cortical neurons' responses. *Science* 224:1260–1262.

Wallach, H. 1976. On perceived identity: 1. The direction of motion of straight lines. In *On Perception*, ed. H. Wallach. New York: Quadrangle.

Wallach, H., O'Connell, D. N. 1953. The kinetic depth effect. *J. Exp. Psych.* 45:205–217.

Wallach, H., Weisz, A., Adams, P. A. 1956. Circles and derived figures in rotation. *Am. J. Psych.* 69:48–59.

Watson, B. W., Nachmias, J. 1977. Patterns of temporal interaction in the detection of gratings. *Vis. Res.* 17:893–902.

Watt, R. J., Morgan, M. J. 1983. The recognition and representation of edge blur: evidence for spatial primitives in human vision. *Vis. Res.* 23(12):1465–1477.

Watt, R. J., Morgan, M. J. 1984. Spatial filters and the localization of luminance changes in human vision. *Vis. Res.* 24(10):1387–1397.

Waxman, A. M., Ullman, S. 1985. Surface structure and three–dimensional motion from image flow kinematics. *J. Robotics Res.* 4:72–94.

Waxman, A. M., Wohn, K. 1987. Image flow theory: A framework for 3–D inference from time–varying imagery. In *Advances in Computer Vision*, ed. C. Brown, New Jersey: Erlbaum *(in press)*.

Westheimer, G., McKee, S. P. 1978. Stereoscopic acuity for moving retinal images. *J. Opt. Soc. Amer.* 68:450–455.

Westheimer, G., McKee, S. P. 1980. Stereoscopic acuity with defocused and spatially filtered images. *J. Opt. Soc. Amer.* 70:772–778.

White, B. W., Mueser, G. E. 1960. Accuracy in reconstructing the arrangement of elements generating kinetic depth displays. *J. Exp. Psychol.* 60:1–11.

Willshaw, D. J., Buneman, O. P., Longuet–Higgins, H. C. 1969. Non–holographic associative memory. *Nature* 222:960–962.

Wilson, H. R., Bergen, J. R. 1979. A four mechanism model for threshold spatial vision. *Vis. Res.* 19:19–32.

Witkin, A. P. 1983. Scale space filtering. In: *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms,* edited by Fischler, M. A., and Firschein, O., Los Altos, CA: Morgan Kaufman, 329–332.

Yasumoto, Y., Medioni, G. 1985. Experiments in estimation of 3–D motion parameters from a sequence of image frames. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, IEEE: New York, 89–94.

Zucker, S. W. 1987. Early orientation selection: Tangent fields and the dimensionality of their support. In: *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms,* edited by Fischler, M. A., and Firschein, O., Los Altos, CA: Morgan Kaufman, 333–347.

This blank page was inserted to preserve pagination.

CS-TR Scanning Project
Document Control Form

Date : 4/27/95

Report # AIM-1038

Each of the following should be identified by a checkmark:
Originating Department:

☒ Artificial Intellegence Laboratory (AI)
☐ Laboratory for Computer Science (LCS)

Document Type:

☐ Technical Report (TR) ☒ Technical Memo (TM)
☐ Other:_____

Document Information

Number of pages: 51 (57 - IMAGES)
Not to include DOD forms, printer instructions, etc... original pages only.

Originals are:
☒ Single-sided or
☐ Double-sided

Intended to be printed as :
☐ Single-sided or
☒ Double-sided

Print type:
☐ Typewriter ☐ Offset Press ☒ Laser Print
☐ InkJet Printer ☐ Unknown ☐ Other:_____

Check each if included with document:

☒ DOD Form (2) ☐ Funding Agent Form ☐ Cover Page
☐ Spine ☐ Printers Notes ☐ Photo negatives
☐ Other: _____

Page Data:

Blank Pages (by page number):_____

Photographs/Tonal Material (by page number): 5, 10, 13

Other (note description/page number):

	Description :	Page Number:
(A)	IMAGE MAP (1) UNNUMBERED TITLE PAGE	
	(2-51) PAGES #'ED 1-50	
	(52-54) SCANCONTROL, DOD(2)	
	(55-57) TARGETS (3)	
(B)	CUT+PASTE FIGS ON PAGES 5, 7, 9, 10, 11, 13, 14, 22, 29-32,	

Scanning Agent Signoff:

Date Received: 4/27/95 Date Scanned: 5/4/95 Date Returned: 5/4/95

Scanning Agent Signature: Michael W. Cash

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

completed 8-26-88

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AI Memo 1038	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER: AD-A195930	
4. TITLE (and Subtitle): The Computational Study of Vision	5. TYPE OF REPORT & PERIOD COVERED: memorandum
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Ellen C. Hildrith and Shimon Ullman	8. CONTRACT OR GRANT NUMBER(s): N00014-85-K-0124
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA 02139	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Advanced Research Projects Agency, 1400 Wilson Blvd., Arlington, VA 22209	12. REPORT DATE: April 1988
	13. NUMBER OF PAGES: 50
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Office of Naval Research, Information Systems, Arlington, VA 22217	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution is unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

computer vision
human vision
binocular stereo
motion analysis
object recognition

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Abstract: Through vision, we derive a rich understanding of what is in the world, where objects are located, and how they are changing with time. Because we obtain this understanding immediately, effortlessly, and without conscious introspection, we can be deceived into thinking that vision should therefore be a fairly simple task to perform. The computational approach to the study of vision inquires directly into the sort of information processing needed to extract important information from the changing visual image – information such

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Block 20 cont.

as the three-dimensional (3-D) structure and movement of objects in the scene, or the color and texture of object surfaces. An important contribution that computational studies have made is to show how difficult vision is to perform, and how complex are the processes needed to perform visual tasks successfully. This article reviews some computational studies of vision, focusing on edge detection, binocular stereo, motion analysis, intermediate vision and object recognition.

Scanning Agent Identification Target

Scanning of this document was supported in part by the **Corporation for National Research Initiatives,** using funds from the **Advanced Research Projects Agency** of the **United states Government** under Grant: **MDA972-92-J1029.**

The scanning agent for this project was the **Document Services** department of the **M.I.T Libraries.** Technical support for this project was also provided by the **M.I.T. Laboratory for Computer Sciences.**